IDENTIFYING GENETIC VARIATION CONTRIBUTING TO KERATOCONUS

by

Sionne Edie Marguerite Lucas

BMedSci, BMedRes (Hons)

Menzies Institute for Medical Research | College of Health and Medicine

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (Medical Studies)

University of Tasmania, December, 2018

DECLARATION OF ORIGINALITY

This thesis contains no material which has been accepted for a degree or diploma by the University or any other institution, except by way of background information and duly acknowledged in the thesis, and to the best of my knowledge and belief no material previously published or written by another person except where due acknowledgement is made in the text of the thesis, nor does the thesis contain any material that infringes copyright.

Sionne Lucas

18/12/2018

AUTHORITY OF ACCESS STATEMENT

The publishers of the papers comprising Chapter 3 hold the copyright for that content, and access to the material should be sought from the respective journals. The remaining non-published content of the thesis may be made available for loan and limited copying and communication in accordance with the *Copyright Act 1968*.

Sionne Lucas

18/12/2018

STATEMENT OF ETHICAL CONDUCT

The research associated with this thesis abides by the international and Australian codes on human and animal experimentation, the guidelines by the Australian Government's Office of the Gene Technology Regulator and the rulings of the Safety, Ethics and Institutional Biosafety Committees of the University.

Sionne Lucas

18/12/2018

STATEMENT OF CO-AUTHORSHIP

Sionne E. M. Lucas has incorporated versions of two first-author papers, Lucas *et al.* (2017)[1] and Lucas *et al.* (2018),[2] into Chapter 3. The contribution of co-authors for each paper is presented below.

Lucas SEM, Zhou T, Blackburn NB, et al. **Rare, potentially pathogenic variants in *ZNF469* are not enriched in keratoconus in a large Australian cohort of European descent**. *Invest Ophthalmol Vis Sci* 2017;58(14):6248-56.

The co-authors and each author's contribution are outlined as follows:

Lucas, S. E. M. was involved the study design, data generation and data curation, she conducted data analysis and laboratory experiments and wrote the manuscript and prepared the tables and figures; Zhou, T. was involved in method development; Blackburn, N. B. was involved in method development, conducted preliminary data analysis and was involved in manuscript revision; Mills, R. A. conducted clinical examinations and was involved in patient recruitment; Ellis, J. was involved in preliminary data analysis; Leo, P. was involved in preliminary data analysis; Souzeau, E. was involved in patient recruitment, data curation and manuscript revision; Ridge, B. was involved in patient recruitment and data curation; Charlesworth, J. C. was involved in study design and manuscript revision; Brown, M. A. was involved in recruitment and data curation; Lindsay, R. conducted clinical examinations and was involved in patient recruitment; Craig, J. E. conducted clinical examinations and was involved in patient recruitment; Burdon, K. P. was involved in study design, funding acquisition, methodology development and manuscript revision.

Lucas SEM, Zhou T, Blackburn NB, et al. **Rare, potentially pathogenic variants in 21 keratoconus candidate genes are not enriched in cases in a large Australian cohort of European descent**. *PLoS One* 2018;13(6):e0199178.

The co-authors and each author's contribution are outlined as follows:

Lucas, S. E. M. was involved in data generation, data curation, and study design, conducted data analysis and laboratory experiments for variant validation, and wrote the manuscript and prepared the tables and figures; Zhou, T. was involved in method development; Blackburn, N. B. was involved in method development, conducted preliminary data analysis and was involved in manuscript revision; Mills, R. A. conducted clinical examinations and was involved in patient recruitment; Ellis, J. was involved in preliminary data analysis; Leo, P. was involved in preliminary data analysis; Souzeau, E.

was involved in patient recruitment, data curation and manuscript revision; Ridge, B. was involved in patient recruitment and data curation; Charlesworth, J. C. was involved in study design and manuscript revision; Lindsay, R. conducted clinical examinations and was involved in patient recruitment; Craig, J. E. conducted clinical examinations and was involved in patient recruitment; Burdon, K. P. was involved in study design, funding acquisition, methodology development and manuscript revision.

Prof. Kathryn Burdon
Supervisor
Menzies Institute for Medical Research
University of Tasmania
18/12/2018

Prof. Alison Venn
Director
Menzies Institute for Medical Research
University of Tasmania
08/05/2019

ACKNOWLEDGEMENTS

As anticipated, an almost four-year project generates a long list of ‘thank yous’.

To the Australian Government for my Research Training Program Scholarship and the Pennicott Foundation for the financial support over the last 3 and a half years.

To all of our collaborators, the clinicians and the participants, without which this study would not be possible. A particularly big thank you to both Emmanuelle Souzeau and Sandra Staffieri for being the in-betweens and answering my many questions about the patients’ and their clinical data.

To Kathryn Burdon, my primary supervisor, for being the most incredible mentor, critic, advocate and friend. Thank you for inspiring me every single day.

To Jac Charlesworth, my co-supervisor, for always pushing me to be better and do better. And for reminding me, at potentially the most critical moment, that the most amazing things are made from crushing pressure and the passage of time.

To Pat and JoJo, for our many ‘coffees’, lunches and early morning chats - your support over the past four years has been immeasurable. You’ve made me laugh on even the hardest of days and helped celebrate some of the best. Thank you being my sounding boards, my teachers, my students and such good friends. Don’t forget about me now that I don’t get to sit with the cool kids.

To Bennet for your bioinformatic support, especially your R expertise, and for keeping Sunnydale running smoothly.

To the rest of our CompGen team and 502 office buddies – Ming, Duran, Elise, Kelsie, Alex, Van, Emma and Aparna – for your support over the years, both professionally and socially.

To James Marthick, ‘the boss’, for all of your sequencing expertise and support.

To Nick, for teaching me all the bioinformatics that you could in the first year of my project. I would have been lost without you and such a strong foundation.

To Sian, for our sporadic and always lengthy phone calls filled with commiserations and encouragement. You’ve got this too.

To Hayley, for being the best cheerleader and friend anyone could ask for. Always.

To my parents, for your never-ending support, advice (while it was sometimes unsolicited, in retrospect it was usually pretty accurate), and encouragement.

To Jenna, my wife (I still have to get used to that!), for everything. For putting up with the stress, the non-existent weekends, and most importantly, the hours of dishes you endured towards the end of my PhD. I’m so glad I’ve had you by my side every step of the way.

ABSTRACT

Keratoconus (OMIM 148300) is a complex disease characterised by progressive stromal thinning and conical protrusion of the cornea. These abnormalities usually develop in the second decade of life and can lead to severe visual impairment as a result of high myopia and irregular astigmatism. Due to the early onset of disease, keratoconus affects individuals during their most productive years and quality of life estimates demonstrate a significant decline over time that is disproportional to visual acuity. The global incidence of keratoconus is approximately 1 in 50,000 and the prevalence in Caucasians is reported between 55 and 265 per 100,000 individuals. While keratoconus is a complex disease, it has a strong genetic basis. Familial examples display both autosomal dominant and recessive inheritance patterns and first-degree relatives have at least a 15-fold increased risk of developing keratoconus, compared to the general population. Current treatment options for keratoconus are suboptimal, require invasive surgical procedures and have limited applicability and effectiveness in early-onset and advanced disease. Currently, keratoconus is the second leading indication for corneal transplants in Australia and there is a great need to better understand keratoconus aetiology and pathophysiology to aid early diagnosis and the development of novel treatments to improve patient outcomes. Therefore, the overarching aim of this project was to identify genetic variants involved in keratoconus susceptibility through the assessment of individuals and families in our keratoconus DNA repository. This aim was achieved through three distinct studies: a candidate gene screen; novel variant discovery in families with multiple cases of severe and early-onset keratoconus; and association analysis in a large case-control cohort.

The gene-screen aimed to determine if rare, protein-coding variants predicted to be pathogenic were enriched in in keratoconus-candidate genes in our cohort of keratoconus cases of European descent compared to ethnically matched controls. By combining whole exome-sequencing and targeted gene sequencing, a total of 22 keratoconus candidate genes were assessed in 385 unrelated cases, making this the most comprehensive study of its kind to date. Two control cohorts were used for comparison, including up to 396 population controls (unassessed for eye disease) and 230 individuals without keratoconus. The candidate genes were selected from the literature and included genes near associated loci, genes harbouring putative disease-causing variants in family studies, and genes proposed to play a role in keratoconus based on a known function in the cornea or involvement in a postulated disease process. Specifically, this study examined the frequency of potentially pathogenic variants in *MPDZ*, *RXRA*, *RAB3GAP1*, *FOXO1*, *BANP*, *ZNF469*, *HGF*, *COL5A1*, *IMMP2L*, *FNDC3B*, *NFIB*, *ILRN*, *SLC4A11*, *CAST*, *COL4A3*, *COL4A4*, *TF*, *SOD1*, *VSX1*, *RAD51*, *IL1A* and *IL1B*. For all genes, no difference was observed in the frequency of rare, protein-coding potentially pathogenic variants between the cases and controls, suggesting that rare protein-coding variation in these genes do not play a major role in keratoconus susceptibility.

The second study utilised whole genome sequencing (WGS) to investigate the genetic basis of disease in two families with severe, early-onset keratoconus. The aim of this study was to identify putatively disease-causing variants that segregate with disease in these families. One family was Jordanian, KCNSW01, with eight cases of keratoconus and three unaffected family members across three generations. While the parents in the first generation are both apparently unaffected, the inheritance pattern in the second-, and third-, generation was indicative of autosomal dominant inheritance of keratoconus. The second family, KSA197, was a family of Italian heritage with two affected brothers born to unaffected, second-cousin parents. Both brothers are affected with keratoconus and one has an unaffected child, consistent with autosomal recessive inheritance. WGS was obtained for 11 individuals from KCNSW01 and five family members from KSA197. A subset of ~250,000 single nucleotide polymorphisms (SNPs) across the autosomes were extracted from the WGS and used to conduct linkage analysis in KCNSW01 and homozygosity mapping in KCNSW01.

Two linkage regions with equal maximum logarithm of the odds (LOD) scores of 2.1 were identified in KCNSW01: 17q12 and 20p13-12.2. The disease-associated haplotypes at 17q12 and 20p13-12.2 were inherited from the matriarch and patriarch (both unaffected), respectively. All affected individuals carry both haplotypes, suggesting digenic inheritance of keratoconus in this family. For KSA197, a single homozygous region shared by the two affected brothers was identified at 16p12.1. For both families, variants that segregated with the disease-associated haplotype(s) were extracted and further investigated. No rare protein-coding variants fulfilled the criteria for putatively disease-causing in either family. Non-coding variation that segregated with disease were therefore prioritised based on the minor allele frequencies in the gnomAD database, predictions of deleteriousness and pathogenicity and whether or not they were located in known regulatory regions. One putatively disease-causing variant was identified in KSA197 and a total of 44 were identified KCNSW01, including a compelling variant located in an untranslated region of the spermine oxidase gene (*SMOX*). This novel candidate gene encodes the SMOX protein which plays a role in apoptosis and the cellular response to ultraviolet radiation and oxidative stress. These pathways are biologically-plausible in the keratoconus disease process and therefore the specific variant and the gene should be further investigated.

The overall aim of the third study was to identify putatively functional variants with a role in keratoconus susceptibility. This study was separated into specific aims. Aim 1 was designed to identify novel keratoconus-associated variants in a large case-control study. As keratoconus is a disease in which the cornea progressively thins, it was hypothesised that variants that contribute to central corneal thickness (CCT) in the general population would also contribute to keratoconus risk. Therefore, 72 SNPs known to contribute to CCT were assessed for association in a cohort of 536 keratoconus cases and 2,574 controls of European descent. Five SNPs were significantly associated with keratoconus following correction for multiple testing, including a novel association at rs2268578 in an intronic region of the lumican gene (*LUM*). The remaining four SNPs – rs1536482 and rs3132303 located

between *RXRA* and *COL5A1*, rs2755238 in the second intron of *FOXO1* and rs66720556 between *MPDZ* and *NFIB* – had previously shown either a significant or suggestive association with keratoconus.

Aim 2 focused on identifying the functional variants that underlie the associations at keratoconus-associated loci. Six regions were fine-mapped: *RXRA-COL5A1*, *FOXO1*, *FNDC3B*, *MPDZ-NFIB*, *RAB3GAP1* and the novel locus *LUM*. This analysis was conducted in a cohort of 487 keratoconus cases and 626 unaffected controls with genome-wide genotyping data. To appropriately capture the association peaks during fine-mapping, variants across the surrounding gene were included for intronic loci, whereas the region encompassing the flanking genes were included for intergenic loci. Strong association peaks were observed at all loci except the previously reported *RAB3GAP1* locus, thus this locus was not further analysed. To further assess variation carried on the risk-associated haplotypes at remaining loci, keratoconus patients carrying the risk allele for the top SNP determined in Aim 2 were selected for re-sequencing in Aim 3. A total of 178 cases and 62 controls were re-sequenced across the five loci. Variants at each locus were filtered to identify those that were carried on the risk-associated haplotype; in high LD with the top SNP as measured by D' to ensure the capture of both common and rare variants; and were more common in the cases compared to the controls and all populations available in Genome Aggregation Database (gnomAD). These variants were further prioritised based on deleteriousness/pathogenicity predictions and whether or not the variant was likely to disrupt a regulatory region. This analysis identified putatively functional variants at all five loci, and proposed rs79728429 as a functional variant at the *FOXO1* (rs2721051) locus. From this work, it was further hypothesised that that rs79728429 alters the expression of a novel uncharacterised gene, *AL133318.1*, and that this altered expression in the cornea confers an increased susceptibility to keratoconus at this locus.

This dissertation comprehensively investigated genetic variation in keratoconus susceptibility in a cohort of Australian keratoconus patients of European descent using three distinct studies and methodologies. In the largest study of its kind to date, this project demonstrated that rare coding variation in 22 keratoconus-candidate genes were unlikely to contribute broadly to keratoconus. The family-based study demonstrated strong evidence of digenic inheritance of keratoconus in one family with the discovery of two linkage regions of equal significance (17q12 and 20p13-12.2) and identified a homozygous region on 16p12.1 in a family with recessive disease. Putatively-disease causing variants within these regions were identified and prioritised for further investigation, including an appealing variant located in the 5' UTR of *SMOX*. Finally, this project identified a novel keratoconus-associated locus overlapping *LUM* with rs3759221 as the top SNP. Putatively functional variants were prioritised at five key keratoconus-associated loci, including a compelling novel putatively functional variant (rs79728429) at the *FOXO1* locus. It was further hypothesised that *AL133318.1* is a regulatory target of rs79728429. Taken together, these findings have contributed substantially to the field of keratoconus

genetics, highlighting the limited contribution of rare coding variation and suggesting a substantial role for non-coding variants in disease susceptibility.

TABLE OF CONTENTS

TABLES

FIGURES

ABBREVIATIONS

µm	micrometre
95% CI	95% confidence interval
AAF	alternate allele frequency
AC	the alternate allele count
AL133318.1	an uncharacterized protein [*gene*]
Alt	alternate allele
ANZRAG	Australian and New Zealand Registry of Advanced Glaucoma
AOGC	Anglo-Australasian Osteoporosis Genetics Consortium (AOGC)
BANP	BTG3 associated nuclear protein [*gene*]
BMES	Blue Mountain Eye Study
bp	base pair
C16orf82	chromosome 16 open reading frame 82 [*gene*]
CADD	combined annotation dependent depletion (an algorithm), usually referring to a scaled CADD score.
CAST	calpastatin [*gene*]
CBS	cystathionine-beta-synthase [*gene*]
CCL5	C-C motif chemokine ligand 5 [*gene*]
CCT	central corneal thickness
CCT6B	chaperonin-containing T-complex polypeptide 1 subunit 6B [*gene*]
CEU	European ancestry, usually referring to individuals from Northern or Weston Europe
CHB	Han Chinese
chr	chromosome
CHST6	carbohydrate sulfotransferase 6 [*gene*]
cM	centimorgan
COL4A3	collagen type IV alpha 3 chain [*gene*]

COL4A4	collagen type IV alpha 4 chain [*gene*]
COL5A1	collagen type V alpha 1 chain [*gene*]
CRISPR	clustered regularly interspaced short palindromic repeats
D’	D-prime
dbSNP	Single Nucleotide Polymorphism Database
DMD	bone mineral density
DNA	deoxyribonucleic acid
DOCK9	dedicator of cytokinesis 9 [*gene*]
dsDNA	double stranded DNA
ENCODE	The Encyclopedia of DNA Elements, a public research project
eQTL	expression quantitative trait loci
ExAC NFE	the non-Finnish European population of the Exome Aggregation Consortium database
ExAC	The Exome Aggregation Consortium database
FATHMM	Functional Analysis through Hidden Markov Models (an algorithm)
FIN	Finnish
FLG	filaggrin [*gene*]
FNDC3B	fibronectin type III domain containing 3B [*gene*]
FOXO1	forkhead box O1 [*gene*]
FS1	filtering strategy 1
FS2	Filtering Strategy 2
GBR	Great Brittan, usually referring to individuals from England and Scotland
GENCODE	a genome research project, part of the ENCODE project
gnomAD	The Genome Aggregation database
GTEx	Genotype-Tissue Expression Project
GWAS	genome-wide association study
H1-hESC	H1 human embryonic stem cell line

HAO1	hydroxyacid oxidase 1 [*gene*]
HapMap3	the International HapMap Project Phase III
HET	heterozygous
hg19	human reference genome version 19
HGF	human growth factor [*gene*]
HMM	Hidden Markov Model
HOM	homozygous
HSMM	Human skeletal muscle myoblasts
HUVEC	Human umbilical vein endothelial cells
IBD	identity-by-descent
ID	identifier
IL-10	interleukin 10
IL1A	interleukin 1 alpha [*gene*]
IL1B	interleukin 1 beta [*gene*]
IL1RN	interleukin 1 receptor antagonist [*gene*]
IMMPL2	inner mitochondrial membrane peptidase subunit 2 [*gene*]
Indels	insertions and deletions, usually referring to these two types of variants collectively
IPA	Ingenuity Pathways Analysis
JPT	Japanese
KDM8	lysine demethylase 8 [*gene*]
KIAA0556	an uncharacterised protein [*gene*]
LD	linkage disequilibrium
LE	left eye
LIG3	DNA ligase 3 [*gene*]
LOD	logarithm of the odds
LOX	lysyl oxidase [*gene*]

MAC	minor allele count
MAF	minor allele frequency
MAVS	mitochondrial antiviral signalling protein [*gene*]
max	maximum
Mb	megabase
min	minimum
MIR184	mircoRNA 184 [*gene*]
MIR8062	microRNA 8062 [*gene*]
miRNA	microRNA
Miss	missing, usually referring to a genotype
mm	millimetres
MPDZ	multiple PDZ domain crumbs cell polarity complex component [*gene*]
n	number
NC	non-carriers
NC	non-coing
NF1B	Nuclear factor 1 B-type [*gene*]
NHEK	normal human epidermal keratinocytes
NHLF	normal Human lung fibroblasts
NSFL1C	NSFL1 cofactor [*gene*]
OMIM	Online Mendelian Inheritance in Man (a catalogue of human genes and genetic disorders and traits)
OR	odds ratio
p	p-value
PC1	the first principal direction, in PCA this is the axis along which the samples show the largest variation
PC2	the second principal direction, in PCA this is the axis along which the samples show the second largest variation

PCA	principle components analysis
PDYN-AS1	PDYN antisense RNA 1 [*gene*]
PLCB1	phospholipase C beta 1 [*gene*]
POAG	primary open-angle glaucoma
PolyPhen2	Polymorphism Phenotyping (an algorithm)
PPCD	posterior polymorphous corneal dystrophy
PRNP	prion protein [*gene*]
r^2	r-squared
RAB3GAP1	RAB3 GTPase activating protein catalytic subunit 1 [*gene*]
RAD51	RAD51 recombinase [*gene*]
RE	right eye
Ref	reference, often referring to the reference allele
RNA	ribonucleic acid
rsID	the variant identifier from the Single Nucleotide Polymorphism Database (dbSNP)
RXRA	retinoid X receptor alpha [*gene*]
sd	standard deviation
SIFT	Sorting Tolerant from Intolerant (an algorithm)
SIRPA	signal regulatory protein alpha [*gene*]
SKAT	Sequence Kernel Association Test
SLC4A11	solute carrier family 4 member 11 [*gene*]
SMOX	spermine oxidase [*gene*]
SNP	single nucleotide polymorphism
SOD1	superoxide dismutase [*gene*]
STK35	serine/threonine kinase 35 [*gene*]
STR	short tandem repeats
TF	transferrin [*gene*]

TGFB1	transforming growth factor beta 1 [*gene*]
UCSC	University of California Santa Cruz
USA	United States of America
UTR	untranslated regions
VCF	variant call format
VPS13D	vacuolar protein sorting 13 homolog D [*gene*]
VSX1	visual system homeobox 1 [*gene*]
WES	whole exome sequencing
WGS	whole genome sequencing
WNT10A	Wnt family member 10A [*gene*]
WT	wild type, usually referring to wild type alleles which is synonymous with reference allele
YRI	Yoruba
ZEB1	zinc finger E-box binding homeobox 1 [*gene*]
ZNF469	zinc finger protein 469 [*gene*]
ZNF830	zinc finger protein 830 [*gene*]

CHAPTER 1: INTRODUCTION

1.1 AN INTRODUCTION TO KERATOCONUS

Keratoconus (OMIM 148300) is a disease characterised by progressive thinning and protrusion of the cornea at the front of the eye, resulting in severe visual impairment. Reports place the prevalence of keratoconus between 17 and 3300 per 100,000[3-14] and the incidence from 1.3 to 32.3 per 100,000 per year.[4-6, 15-18] The prevalence and incidence rates vary greatly between studies, methods of diagnosis, population and geographic location, with markedly higher statistics in Asians[19] and, anecdotally, in Polynesians.[20] The most commonly reported prevalence in Caucasians, which this dissertation will focus on, is 54.5 per 100,000 with an incidence of 2.0 per 100,000 per year.[4] Keratoconus is typically bilateral, although the development and progression may be asymmetrical.[21] Diagnosis is generally made around puberty or during early adulthood, however in the early stages of disease, keratoconus is difficult to differentiate from regular refractive errors and may be misdiagnosed.[22, 23] Furthermore, diagnosis before the age of 19 is associated with severe disease characterised by a faster rate of progression as well as diagnosis at a more advanced stage, compared to adults.[24] However, the only current treatment with the potential to inhibit keratoconus progression, a surgical procedure known as corneal collagen cross linking, cannot be used in severe disease as a minimum corneal thickness is required to ensure safety.[25] This procedure requires early diagnosis of keratoconus and may exclude patients most in need of treatment. Unfortunately, all other current treatments merely manage symptoms and ultimately one in five keratoconus patients will require corneal transplants.[26] Therefore, there is a great need for the development of novel treatments as well as a better understanding of the aetiology and pathophysiology of disease to aid early diagnosis.

While our understanding of the aetiology of keratoconus is limited, it is clear that both environmental and genetic factors involved in disease susceptibility and development. Furthermore, the aetiology is heterogeneous, with the relative contributions of genetic and environmental factors differing between individuals. At least one case in the literature seems to have resulted entirely from traumatic injury,[27] however, there are also cases demonstrating clear autosomal dominant patterns of inheritance.[28, 29] Therefore, it is hypothesised that in the majority of cases environmental factors are required to trigger keratoconus development in genetically predisposed individuals. As extensively reviewed by Gordon-Shanng and colleagues,[30] environmental risk factors include eye rubbing, atopy (hypersensitivity reactions in the form of eczema, asthma and allergy), ultraviolet light exposure and geographic location, however, largely the genetic risk factors involved in keratoconus predisposition have not yet been elucidated. Identifying specific genetic factors involved in keratoconus susceptibility and pathogenesis would allow for improved genetic counselling, earlier diagnosis as well as pave the way for the

development of biomarkers, novel therapies and management strategies, which taken together would greatly improve patient's quality of life.

1.2 THE HUMAN EYE AND VISION

The eye is a complex organ that receives input for the visual sensory system. The eye responds to light stimuli and relays this information to the brain, where visual perception occurs. For good vision, the coordination of highly specialised structures within the eye is critical and involves the transparency of tissues, the regulation of light intensity, and the coordinated refraction of light onto the photosensitive cells of the retina. Upon light stimulation, photosensitive cells produce action potentials which are conducted along the optic nerve from the eye to the brain.

1.2.1 Basic anatomy and physiology of the human eye

The eye is essentially an irregular sphere, comprising of layers of tissues surrounding a largely fluid-filled space (Figure 1.1). The surface of the eye consists of the sclera and the cornea, both of which are collagen-rich connective tissues that provide structural integrity and protect the intraocular structures. The sclera is a tough, opaque structure that forms the majority of the surface area of the eye and is continuous with the cornea. The cornea is a transparent tissue that forms a dome-shaped at the front of the eye. In addition to its protective role, the cornea allows the passage of light into the eye and is responsible for two-thirds of the eye's refractive power. Behind the cornea the iris is clearly visible. The iris is a highly pigmented ring-shaped tissue that acts like a diaphragm to regulate the amount of light that can pass through the pupil. The crystalline lens is suspended behind the iris and divides the eye into the anterior and posterior chambers. Like the cornea, the lens is transparent and is involved in the focusing of light onto the retina. In contrast to the cornea which has a constant refractive power, the refractive power of the lens can be altered in a process known as accommodation. The retina lines the internal surface at back of the eye and consists of layers of neural cells as well as photoreceptors.

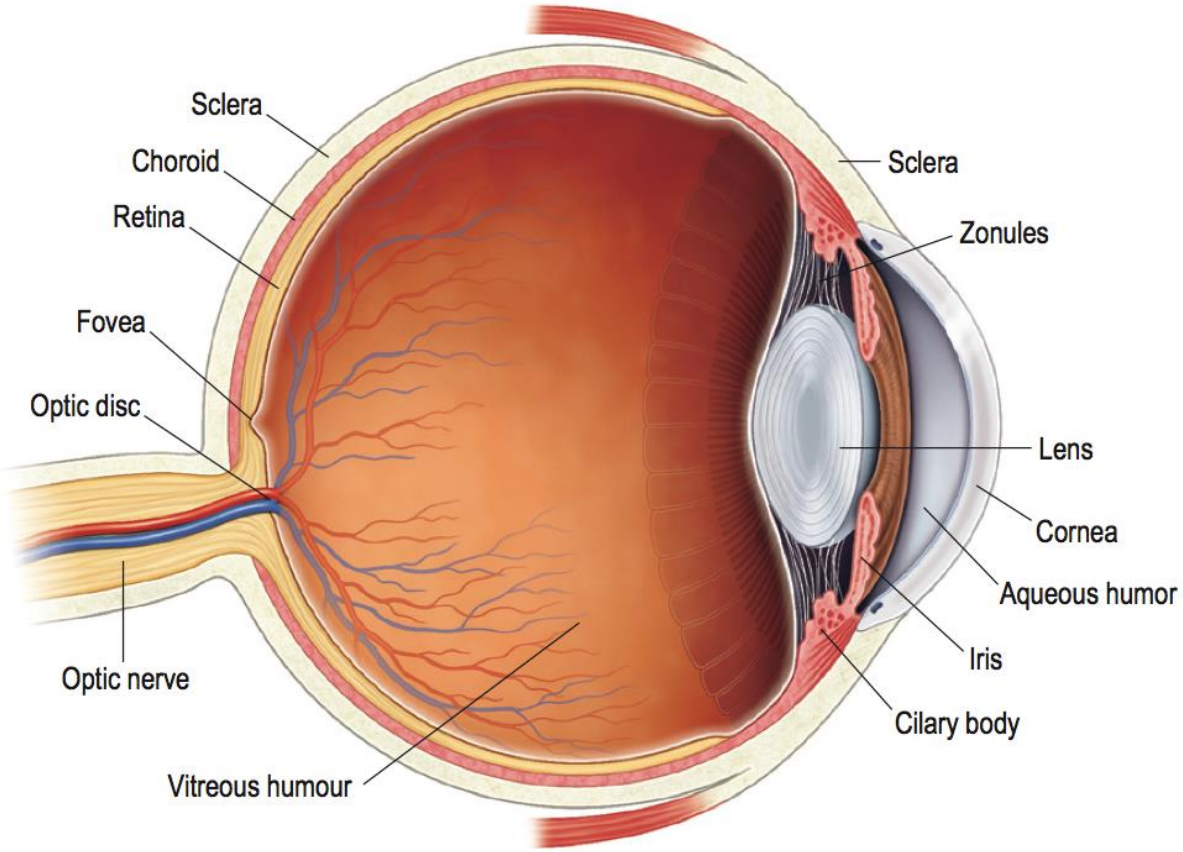


Figure 1.1 – A diagram of the cross section of a human eye.
Source: Levin, LA, Nilsson, SFE, Ver Hoeve, J, Wu, SM, Kaufman, PL & Alm, A (eds) 2011, Adler's Physiology of the Eye, 11th edn, Saunders Elsevier, Sydney.

1.2.2 The cornea

The cornea is the transparent portion of the outermost tunic of the eye. On a gross level, the cornea is aspherical with an apex where the curvature is the steepest. Corneal curvature gradually decreases towards the scleral junction. The thinnest region of the cornea is at the centre with corneal thickness increasing towards the sclera. Central corneal thickness (CCT) is a quantitative trait with normal distribution in Caucasians around a mean of 0.536 mm with the normal range between 0.473 and 0.597 mm.[31] The cornea is avascular as blood vessels would inhibit the transmission of light and thus the cells of the cornea are nourished by the surrounding fluids: the aqueous humour contained within the anterior chamber, or the tear film that coats the anterior surface.

On a finer scale, the cornea is comprised of five discrete layers (from anterior to posterior): the epithelium, Bowman's layer, the stroma, Descemet's membrane and the endothelium (Figure 1.2). The epithelium is a five to seven cell layer that forms the outer surface of the cornea. The primary function of the epithelium is to provide a smooth refractive surface, but it also forms a protective barrier against and pathogens and fluid loss. Bowman's layer forms a membrane of woven collagen fibrils between the epithelium and the stroma that functions largely to support the shape of the cornea, as well as provide

protection for the deeper layers. The central layer, the stroma, makes up the 90% of the corneal thickness and is integral for the shape, structural integrity and transparency of the cornea. The stroma consists of a collagen-rich extracellular matrix interspersed with supporting keratocytes. Keratocytes have a characteristic stellate-like morphology, resulting from numerous lamellapodia that extend from the compact cell body, allowing for cell-to-cell communication, whilst minimising light scattering.[32] These cells are essential for the production and maintenance of the surrounding extracellular matrix, including the excretion of collagens and proteogyclycans.[33] Collagen type I is the primary collagen found throughout the stromal layer.[34] The second most abundant collagen is type V and many other collagens are present in lesser quantities.[32, 34] Along with proteoglycans, these collagens form water-soluble fibrils, which are regularly packed and are arranged into lamellae. In the anterior third of the stroma, these lamellae are interlaced in three dimensions, and this is thought to contribute to corneal rigidity.[35] The lamellae located in the posterior portion of the stroma form orthogonal layers, which is important for transparency.[36] Beneath the stroma, the Descemet's membrane forms an elastic collagen-rich lattice structure that supports the underlying endothelium. The endothelium is a single-cell layer and is primarily involved in the regulation of the fluid levels in the cornea. For corneal transparency, the stroma must exist in a state of relative dehydration.[32] This is achieved via the active transport of ions into the aqueous humour by endothelial cells, creating an osmotic gradient and draws fluid out of the stroma.[32]

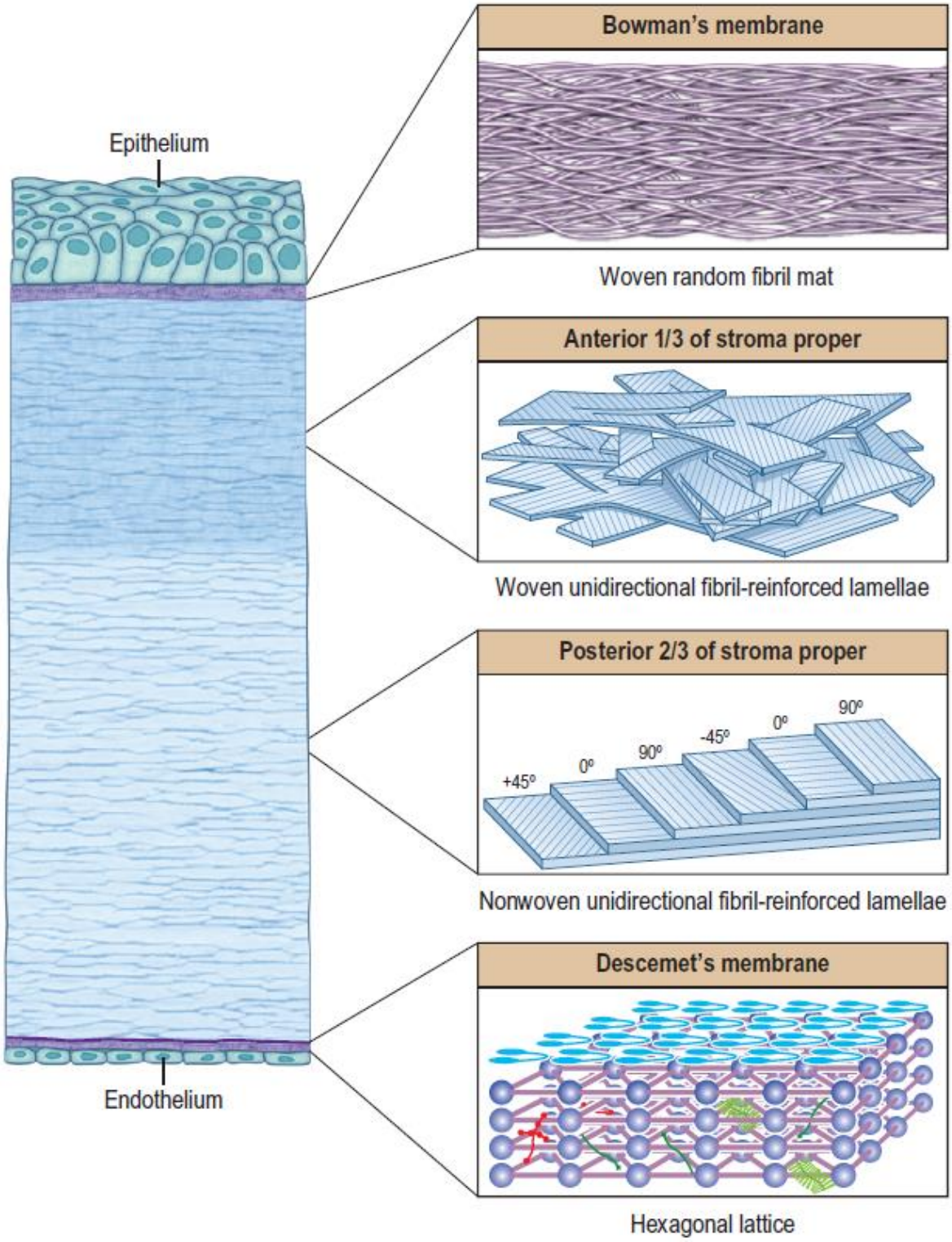


Figure 1.2 – A diagram of layers of the mature cornea.
Adapted from: Levin, LA, Nilsson, SFE, Ver Hoeve, J, Wu, SM, Kaufman, PL & Alm, A (eds) 2011, Adler's Physiology of the Eye, 11th edn, Saunders Elsevier, Sydney.

1.3 KERATOCONUS

1.3.1 Signs and symptoms

Symptoms are subtle in early in keratoconus and patients may simply notice blurred vision. As the disease progresses the cornea increasingly protrudes, forming a conical shape, and visual acuity diminishes due to the occurrence of high myopia and irregular astigmatism. High myopia is a severe form of short-sightedness and is a characteristic of keratoconus due to the increased axial length of the eye, resulting from the conical protrusion of the cornea.[37] Irregular astigmatism is also characteristic of keratoconus and results in multiple focal points and distorted vision due to the irregular surface and shape of the cornea.

Clinically, the first conclusive sign of keratoconus is steepening of the cornea,[38] however, corneal thinning commonly precedes this.[39] In keratoconus, the degree of corneal thinning can vary greatly between individuals, however, a meta-analysis of CCT in keratoconus patients showed a mean of 0.434 mm across 12 studies.[31] In addition, distortion of keratometric images and abnormal light reflexes, such as scissoring of the retinoscopic reflex and oil droplet sign of Charleux, are early signs of keratoconus.[30, 40] As the disease progresses pigmented iron deposits, known as Fleischer's ring, are commonly observed around the base of the developing cone.[39] Fine vertical stress lines, known as Vogt's striae, resulting from the compression of the Descemet's membrane are also a sign of moderate keratoconus. Advance stages of keratoconus are characterised by V-shaped deformation of the lower eyelid during downward gaze known as Munson's sign, Rizzuti's sign (an abnormal light reflection), scaring of the cornea, and corneal hydrops.[39] Corneal hydrops is a potentially-blinding complication of keratoconus and is caused by breaks in the Descemet's membrane that results in oedema of the stroma and loss of corneal transperancy.[39] Corneal hydrops presents with severe pain, corneal opacity and photophobia and requires immediate corneal transplantation.[41]

The gold standard technique for diagnosis and monitoring of keratoconus progression is corneal topography, a computer-assisted method for mapping and analysing the corneal surface and curvature.[42] This highly sensitive method has allowed for the development of a number of highly specific indices for the diagnosis of keratoconus[43, 44] and also allows for the identification of subtle manifestations that don't meet the strict criteria for keratoconus, commonly referred to in the literature as forme fruste keratoconus, keratoconus suspect or subclinical keratoconus.[22, 23] Forme fruste keratoconus is important to diagnose and monitor as it may progress and develop into true keratoconus later in life. The diagnosis of forme fruste keratoconus is particularly critical in individuals undergoing refractive surgery as the procedure weakens the cornea and may trigger keratoconus devleopment.[45, 46]

1.3.2 Treatments and management

In early stages of disease, the refractive errors associated with keratoconus can be corrected with glasses, however as the disease progresses, contact lenses are required.[47, 48] Contact lens fitting can be a complex and it may be difficult to achieve adequate visual acuity with comfort and without compromising corneal health.[47, 48] A range of contact lenses may be used depending on the severity of disease and comfort of the patient including: soft, ridged gas permeable, scleral and hybrid lenses.[47, 48] In addition, piggyback lenses is common and involves wearing two lenses on one eye, generally a soft contact lens underneath a gas permeable lens.[47, 48] However, as the disease progresses contact lens intolerance, corneal abrasions and scaring may develop and corneal transplantation may be required.[47] Currently, keratoconus is the second leading indication for corneal grafting in Australia. Corneal transplantation is a surgical procedure in which a central portion of the corneal is replaced with a healthy graft from a deceased donor. However, graft survival time is limited and due to the early age of onset of disease, keratoconus patients may require multiple grafts throughout their lifetimes.[49-53] Like any surgical procedure, corneal transplantation is associated with a number of well documented complications, however, regrafting is associated with increased complications, decreased graft survival and poorer visual acuity.[49, 54-56]

In recent years the development of corneal collagen cross-linking, a surgical procedure that induces covalent bonds between the collagen fibres of the corneal stroma, has shown promising results in the inhibition of keratoconus progression. The procedure involves the removal of the epithelium followed by the addition of riboflavin (vitamin B_2) and irradiation with long wave ultraviolet light (UVA, 370nm).[57] This results in increased corneal rigidity through a process known as photopolymerisation.[57] However, corneal collagen cross-linking can only be conducted on corneas with a minimum CCT of 0.400 mm to ensure protection of the sensitive endothelium and therefore cannot be used to treat well progressed cases of keratoconus.[58] Thus, corneal collagen cross-linking requires timely diagnosis of keratoconus. As keratoconus tends to be diagnosed at a later stage and progress faster in paediatric cases,[24] this may exclude patients most in need of this treatment. Furthermore, as reviewed by Mastropasqua,[25] corneal collagen cross-linking fails to inhibit disease progression in up to a third of cases. A better understanding the underlying genetic risk factors and pathogenesis of disease would lead to the development of novel treatments to inhibit keratoconus progression and potentially identify biomarkers to aid early diagnosis, which together would greatly improve patient outcomes.

1.4 KERATOCONUS GENETICS

There is strong evidence for the role of genetic factors in the aetiology of keratoconus. Up to 27.9% of patients report a family history of disease, implicating the role of genetic factors in these cases.[4, 59-61] It is however widely accepted that sporadic cases of keratoconus are a result of environmental triggers in genetically predisposed individuals. Evidence for this includes the bilateral nature of the disease, the 15 to 67 fold higher prevalence in first-degree family members of keratoconus patients compared to the general population[62] and the identification of undiagnosed keratoconus and forme fruste keratoconus in family members of keratoconus cases that were previously believed to be sporadic.[62, 63] Taken together, this suggests that genetic factors contribute to keratoconus susceptibility in a large proportion of cases.

1.4.1 Linkage studies

To date, 16 linkage studies for keratoconus have been published (Table 1.1). These linkage regions map to 12 of the human autosomes, demonstrating the heterogeneity of genetic factors involved in keratoconus. The identification of the causal variants in these linkage regions has been slow and difficult due to the complex nature of keratoconus. Linkage studies rely on highly penetrant traits and clearly defined phenotypes and therefore reduced penetrance, the occurrence of phenocopies and heterogeneous phenotypes all contribute to the difficulty of identifying specific genetic variants in families with keratoconus. Despite this, putatively causal variants have been proposed in four of the linkage studies, implicating the role of genetic variation within the associated genes, *DOCK9*, *MIR184*, *SLC4A11* and *IL1RN*, in keratoconus pathogenesis.[64-66] While variation in these genes have not been widely investigated, preliminary findings suggest that these genes are involved in the pathogenesis of keratoconus, but they may only account for disease in the family in which the gene was first implicated,[67, 68] or in the case of *MIR184*, in very few additional cases.[69, 70] Genetic screening of unrelated keratoconus patients is required to confirm and determine the extent to which these genes contribute to keratoconus predisposition. In addition, further analysis of the genetic variation within families in which causal variants have not yet been elucidated presents an opportunity to identify novel genes involved in keratoconus susceptibility.

Table 1.1 – Statistically significant loci from keratoconus linkage studies.

Locus	Statistic	Cohort type	Population	Number of families	Number of individuals (affected)	Identified variants	Ref
1p36.23-36.21*	LOD = 3.4; NPL = 7.8 (P = 0.00024)	Large family	Australian (Caucasian)	1	19 (9)		[71]
2p24	HLOD = 5.13	Small families	Caucasian, Arab and Caribbean African	28	253 (112)		[72]
2q13-q14.3[#]	NPL = 2.4	Large Family	Ecuadorian	1	21 (9)	*IL1RN* c.214+242C>T	[66]
3p14-q13	LOD = 3.09	Large family	Italian	1	21 (11)		[28]
5q15-q21.1	LOD = 3.48	Large family	American (Caucasian)	2 (intermarried)	27 (14)		[73, 74]
8q13.1-q21.11*	LOD = 3.4; NPL = 7.8 (P = 0.00024)	Large family	Australian (Caucasian)	1	19 (9)		[71]
9p21 (34 cM)	LOD = 3.8; NPL = 5.55 (P = <0.001)	Families with affected sibling pairs	Hispanic	17 (sibling pair families)	93 (≥ 17 sibling pairs)		[75]

Locus	Statistic	Cohort type	Population	Number of families	Number of individuals (affected)	Identified variants	Ref
9q34 (159cM)	LOD = 4.5	Families with affected sibling pairs	Caucasian and Hispanic	67 (sibling pair families)	351 (110 sibling pairs)		[75]
13q32	LOD = 4.1	Small and large families	Ecuadorian	18	143 (76)	*DOCK9* c.226A>C	[29, 64]
14q24.3	LOD = 3.58	Small and large families	Caucasian, Iranian, Indian and Pakistani	6	36 (21)		[76]
15q22.33-q24.2	LOD = 8.13	Large family	Northern Irish	1	30 (16)	*MIR184* r.57C>U	[65, 77]
16q22.3-q23.1	LOD = 4.10; NPL = 3.27 (P = 0.00006)	Small families	Finnish	20	76 (42)		[78]
17p13	LOD = 3.21	Large family	Pakistani	1 (consanguineous)	18 (4)		[79]
17q24 (86cM)	LOD = 3.9; NPL = 3.32 (P = <0.001)	Families with affected sibling pairs	Hispanic	17 (sibling pair families)	93 (≥ 17 sibling pairs)		[75]

Locus	Statistic	Cohort type	Population	Number of families	Number of individuals (affected)	Identified variants	Ref
20p13-p12.2[#]	NPL = 2.7	Large Family	Ecuadorian	1	21 (9)	*SLC4A11* c.2558+149_2558+203del154	[66]
20q12	P = $2.1x10^{-5}$	Identity-by-descent	Northern Tasmanian (Caucasian)	1 six-generation family identified (4 of 8 cases were distantly related)	8 (8)		[80]

* refers to digenic inheritance.
[#] indicates a suggestive locus were a putatively causative variant was identified.
LOD = logarithm of the odds.
HLOD = heterogeneity LOD.
NPL = nonparametric LOD.
P = P-value.
Ref = reference.
Adapted from Lucas, SEM.[81]

1.4.2 Genome-wide association studies

Association studies aim to identify variants that are more commonly observed in cases compared to controls. Recently, high throughput DNA genotyping technology has allowed for genome-wide association studies (GWAS) which detect associations throughout the genome, without *a priori* hypotheses. Genome-wide association studies (GWAS) aim to detect associations between the genotypes at thousands of single nucleotide polymorphisms (SNPs) and complex traits in hundreds of unrelated cases and controls. GWAS are designed to exploit linkage disequilibrium (LD), the non-random association (or co-inheritance) of alleles at different loci, at a population level. LD is broken down by recombination events over generations, and therefore within a given population, LD blocks reflect haplotype blocks interspersed by recombination hotspots.[82] SNPs are selected such that they 'tag' LD blocks and therefore SNPs that meet the stringent significance threshold of 5×10^{-8} identify risk-associated loci for further investigation. An important next step is fine-mapping associated loci, which requires genotyping additional SNPs across the LD block and assessing them for association with the trait. This method aims to confirm the association signal, identify the specific risk-associated haplotype(s) and ultimately is important for elucidating the functional variant(s).[83]

Two significant, or highly suggestive, keratoconus-associated loci have been identified through GWAS for keratoconus: rs3735520 and rs17501108 on 7q21.11 (upstream of *HGF*)[84] and rs4954218 at 2q21.3 (upstream of *RAB3GAP1*).[85, 86] In addition, Lu and colleagues[87] elegantly demonstrated that two loci associated with the quantitative trait, CCT, were also associated with keratoconus: rs2721051 on 13q14.11 (downstream of *FOXO1*) and rs4894535 on 3q26.31 (in an intronic region of *FNDC3B*). This study genotyped 26 CCT-associated loci in keratoconus patients based on the hypothesis that genetic factors involved in CCT may also contribute to keratoconus as reduced CCT is a feature of the disease. Four additional CCT-associated loci showed a suggestive association with keratoconus in this study: rs1536482 and rs7044529 on 9q34.3 (between *RXRA* and *COL5A1* and in an intron of *COL5A1*, respectively), rs1324183 on 9p23 (between *MPDZ* and *NF1B*) and rs9938149 on 16q24.2 (between *BANP* and *ZNF469*). The *MPDZ-NFIB* locus later reached genome-wide significance following replication and meta-analysis.[88] Together these studies identified four susceptibility loci for keratoconus, implicated a further four loci with highly suggestive association with keratoconus, and demonstrated the effectiveness of assessing CCT-associated SNPs in keratoconus patients to identify novel disease-associated loci.

1.4.3 Functional gene candidates

There have been many candidate genes suggested to be involved in keratoconus based on the known function or expression in the cornea: *CAST, COL4A3, COL4A4, TF, SOD1, VSX1, RAD51, IL1B, IL1A, FLG, ZEB1, LOX*. Of these genes, *VSX1,* [67, 89-102] *SOD1*[95, 96, 102-106] and *LOX*[96, 107-110] have been screened in a number of relatively small case-control studies, however, the findings have been inconsistent and

therefore further analysis is warranted. The remaining functional candidates have not been widely assessed in keratoconus and thus large case-control studies are required to confirm the role of these genes in keratoconus susceptibility.

1.4.4 The future of keratoconus genetics

While progress in the field of keratoconus genetics has been slow, currently 30 candidate genes have been proposed to play a role in disease susceptibility (Table 1.2). To date, four keratoconus-associated loci have reached genome-wide significance and another four show a suggestive association. Nearby genes have been hypothesised to play a role keratoconus susceptibility, however, the keratoconus-associated loci require fine-mapping to identify the functional variants. Additionally, the success of assessing CCT-associated loci in keratoconus patients warrants the analysis of additional CCT-associated loci in keratoconus to elucidate novel genetic factors. Moreover, linkage studies have identified four putative susceptibility genes, however the contribution of these genes to keratoconus development more broadly is not yet known and requires further study. The complexity of keratoconus genetics has impeded the identification of casual variants in the remaining linkage studies, although, it is expected that due to the increased accessibility to massively parallel sequencing techniques, such as whole exome sequencing (WES) and whole genome sequencing (WGS), the number of causal variants identified in keratoconus families will increase substantially in coming years. The utility and effectiveness of this technology was recently demonstrated through the identification of a novel potentially-causative variant in the transforming growth factor beta-1 gene (*TGFB1*; c.T1209G) in an Iranian family with keratoconus.[111] This gene had previously been suggested to play a role in keratoconus susceptibility based on the role of the protein in cellular interactions with the extra-cellular matrix in healthy corneas and its decreased expression in keratoconic corneas.[112] This example highlights how an understanding of corneal biology can successfully identify candidate genes and the role of massively parallel sequencing techniques in the assessment of these genes. Furthermore, WES in large cohorts of unrelated keratoconus patients may allow for the identification of additional candidate genes without *a priori* hypotheses. Therefore, it is hoped that massively parallel sequencing techniques will aid family studies, as well as population-based studies, and represent a prosperous new chapter in the field of keratoconus genetics.

Table 1.2 – Candidate genes for keratoconus proposed in the literature.

Gene	Study Type	Why Studied?	Initial Studies
CAST	Case-control	Candidate gene within linkage region	[74, 113]
COL4A4	Case-control	Expression/function in the cornea	[114]
TF	Case-control	Functional candidate	[115]
COL4A3	Case-control	Functional candidate and corneal expression	[114]
ZNF469	GWAS	Proximity to GWAS signal	[87]
MPDZ	GWAS	Proximity to GWAS signal	[87]
RXRA	GWAS	Proximity to GWAS signal	[87]
RAB3GAP1	GWAS	Proximity to GWAS signal	[85, 86]
IL1RN	Linkage	Candidate gene within a linkage region	[66]
FOXO1	GWAS	Proximity to GWAS signal	[87]
BANP	GWAS	Proximity to GWAS signal	[87]
HGF	GWAS	Proximity to GWAS signal	[84]
SOD1	Case-control	Functional candidate	[103]
VSX1	Case-Control	Corneal expression	[89]
COL5A1	GWAS	Proximity to GWAS signal	[87]
RAD51	Case -control	Functional candidate	[116]
IMMPL2	GWAS	Proximity to GWAS signal	[85, 86]
SLC4A11	Linkage	Candidate gene within a linkage region	[66]
VPS13D	Linkage	Candidate gene within a linkage region	[71, 81]
IL1B	Case-control	Functional candidate	[117]
IL1A	Case-control	Functional candidate	[117]
FNDC3B	GWAS	Proximity to GWAS signal	[87]
NFIB	GWAS	Proximity to GWAS signal	[87]
CBS	GWAS	Proximity to GWAS signal	[118]
TGFB1	Case-control	Functional candidate and family study	[111, 112]
FLG	Case-control	Functional candidate	[119]
DOCK9	Linkage	Candidate gene within a linkage region	[29, 64, 120]
MIR184	Linkage	Candidate gene within a linkage region	[65, 77]
ZEB1	Case-control	Functional candidate	[121]
LOX	Linkage	Candidate gene within a linkage region	[75, 107]

1.5 PROJECT SUMMARY

The heterogeneous nature of the genetic basis of keratoconus suggests that keratoconus is in fact a collection of disorders, perhaps caused by defects in various proteins that function within the same biological pathway, all of which result in the same phenotype. The identification of novel candidate genes would aid our understanding of keratoconus pathogenesis, as well as contribute to our understanding of corneal biology. This could lead to the development of novel treatments and the identification of biomarkers to aid early diagnosis to improve patients' lives. Therefore, this project aims to explore the field of keratoconus genetics in a cohort of more than 620 Australian keratoconus patients of European ancestry to identify genetic risk factors involved in keratoconus susceptibility. To achieve this, a combination of methodologies will be applied including: whole-exome and whole-genome sequencing to interrogate genetic variation in familial and severe sporadic cases; targeted sequencing of candidate genes in unrelated cases to investigate genetic variation in disease; as well as genotyping and association analysis of key loci in a large case-control study. Together, these studies will address two distinct but complimentary hypotheses.

1.5.1 Hypotheses and aims

Hypothesis 1: ***Rare, highly penetrant protein-coding variants contribute to keratoconus development.***

Overall aim: **To investigate the role of rare protein-coding variants that are predicted to be damaging (potentially pathogenic variants) in Australian keratoconus patients of European descent.**

More specifically to:

1. Determine if potentially pathogenic variants are enriched in known candidate genes in keratoconus patients compared to controls; and
2. Identify rare, putatively disease-causing variants in families with multiple cases of early-onset or severe keratoconus and a strong Mendelian inheritance pattern of disease.

Hypothesis 2: ***Variants associated with keratoconus indicate haplotypes that harbour functional variants which directly contribute to keratoconus susceptibility.***

Overall aim: **To identify variants that contribute to keratoconus susceptibility in a large cohort of unrelated Australian keratoconus patients with European ancestry**.

More specifically to:

1. Identify novel keratoconus-associated loci by assessing central corneal thickness-associated loci in keratoconus patients and unaffected controls;

2. Fine-map keratoconus-associated loci in a cohort of unrelated keratoconus cases and controls to investigate the extent of the association, identify the top SNP, and select genomic regions for re-sequencing. This aim will assess novel keratoconus-associated loci identified in Aim 1, as well as, published keratoconus-associated loci that have reached genome-wide significance.
3. Identify putatively functional variants underlying keratoconus-associated loci by re-sequencing keratoconus patients carrying the risk-associated alleles. This aim will focus on fine-mapped regions with strong association peaks identified in Aim 2.

CHAPTER 2: GENERAL MATERIALS AND METHODS

2.1 STUDY COHORTS

All investigations adhered to the principles of the Declaration of Helsinki and were approved by the Southern Adelaide Clinical Human Research Committee and the Human Research Ethics Committee Tasmania. All participants gave written informed consent.

2.1.1 Keratoconus patients

Keratoconus patients were recruited through the Flinders Eye Clinic (Adelaide, Australia) by referral from their treating optometrist or ophthalmologist. Additional patients were recruited from across Australia via mail through Keratoconus Australia. This cohort currently consists of approximately 630 patients. All clinical examinations were performed by an experienced ophthalmologist. Individuals were diagnosed with keratoconus if they had videokeratographic features of keratoconus or any of the following signs: conical corneal protrusion, central or paracentral stromal thinning or other distinctive features such as Fleischer's ring, Vogt's striae, epithelial or sub-epithelial scarring, or oil droplet sign and/or scissoring of the retinoscopic reflex. In addition, individuals with a history of corneal transplantation for keratoconus were also classified as cases. Where a family history of disease was reported, both affected and unaffected family members were invited to be involved in the study. Family pedigree figures were generated using Cranefoot[122] (version 3.2.3). Cohorts of unrelated keratoconus patients were used in Chapters 3 and 5, while key families were investigated in Chapter 4.

Peripheral blood samples were collected from the recruited individuals and DNA was extracted using the QiaAmp DNA Maxi Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. For a small number of individuals saliva samples, rather than blood samples, were collected using the Oragene OG-500 DNA collection kit (DNA Genotek Inc., Ottawa, Ontario, Canada) and DNA was extracted with the Prep-It reagent (DNA Genotek Inc.) according to manufacturer's instructions.

2.1.2 Control cohorts

Throughout this dissertation, older cohorts of individuals without clinical signs of keratoconus were selected as controls, wherever possible. As keratoconus generally develops during early adulthood, but can develop at any age, older control cohorts were selected to minimise the risk that these individuals would develop the disease and therefore be misclassified as unaffected individuals. The various control cohorts are outlined in sections 2.1.2.1 through to 2.1.2.4.

2.1.2.1 Population controls from the Anglo-Australasian Osteoporosis Genetic Consortium (AOGC)

This cohort consists of 993 individuals from the Anglo-Australasian Osteoporosis Genetics Consortium (AOGC) were used as a population control cohort in Chapter 3. These individuals were ethnically matched females with moderately high, or low, bone mineral density measurements (1.5<|BMD|<4.0). These individuals were not examined for eye disease and therefore the frequency of keratoconus in this cohort is expected to be equal to the population frequency (approximately 1 in 1,500). This cohort has been described in detail previously.[123]

2.1.2.2 Screened controls from the Australian and New Zealand Registry of Advanced Glaucoma (ANZRAG)

This cohort consists of 230 individuals that were examined for eye disease by experienced ophthalmologists at the Flinders Eye Clinic. While these individuals had no clinical evidence of keratoconus, the majority of these individuals had advanced glaucoma (n=195). The remaining individuals either had no evidence of eye disease (n=22) or were unaffected individuals from families with congenital cataract (n=9) and nanophthalmos (n=4). This cohort was used in Chapter 3.

2.1.2.3 Screened controls from the Blue Mountain Eye Study (BMES)

These individuals were examined by an ophthalmologist and were deemed to be unaffected by keratoconus, although have other eye diseases such as glaucoma and cataract at normal population frequencies. All individuals in this cohort were above the age of 50 at the time of recruitment. This cohort includes 2574 individuals and has previously been described in detail by Mitchell and colleagues.[124] These individuals were used as a screened control cohort in Chapter 5.

2.1.2.4 NSA controls

These individuals were examined for glaucoma, including a family history, but were not examined for other eye diseases, although any obvious eye conditions were noted during the examination. This cohort included 199 individuals who were over age 50 at the time of recruitment and were living in residential care facilities in Adelaide, South Australia. This cohort was used as population controls in Chapter 5.

2.2 MASSIVELY PARALLEL SEQUENCING

Massively parallel sequencing was used throughout this dissertation, with targeted re-sequencing methods used in Chapters 3 and 5 and whole genome sequencing in Chapter 4. Specific details of the data generation and variant calling methods are outlined in the relevant chapters, however, in all

chapters, identified variants were output into variant call format (VCF) files. Methods for the manipulation of these files was consistent across the chapters and are outlined below.

2.2.1 Adding confidence tags to genotypes in VCF files

Confidence tags based on read depth and genotype quality were added to all genotypes within a VCF using the VariantFiltration tool from GATK[125] in R.[126] Variants with a depth of at least 10 and a genotype quality score of at least 20 were tagged as 'high confidence', those with a depth below 10 and a genotype quality score of at least 20 were tagged as 'low coverage', variants with a depth of at least 10 and a genotype quality score below 20 were tagged as 'low quality', and those with a depth below 10 and a quality score below 20 were tagged with 'low confidence'. The script was developed by Dr. Nicholas Blackburn (Menzies Institute for Medical Research, University of Tasmania, TAS, Australia; and South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Brownsville, Texas, USA). An example is outlined in Appendix 1.

2.2.2 Converting variants with low coverage/quality/confidence genotype calls to missing calls in VCF files

Variants with low coverage, low genotype quality or low confidence tags (as outlined in Section 2.2.1) were converted to missing genotype calls using a custom R script. This script was developed in combination with Dr. Nicholas Blackburn and Dr. Bennet McComish (Menzies Institute for Medical Research, University of Tasmania, TAS, Australia) and an example is presented in Appendix 2.

2.2.3 Annotating VCF files with ANNOVAR

ANNOVAR[127] (2017Jun01 version) was used throughout this dissertation to annotate variant call format (VCF) files with variant identification codes (IDs) from the dbSNP[128] 147 variant database, gene annotations from the RefGene[128] database, minor allele frequencies (MAF), as well as, deleteriousness/pathogenicity predictions from *in silico* tools. The Exome Aggregation Consortium database[129] (ExAC), the Genome Aggregation database[129] (gnomAD), Kaviar Genomic Variant Database[130] (Kaviar) and the 1000 Genomes Project[131] (1KGP) were used for MAF annotations. The gnomAD database is the largest database with more than 130,000 individuals across seven defined ethnic populations (additional individuals are groups together in the population group 'Other') and includes data from the 1KGP and ExAC, therefore, the MAF observed in these data was of primary interest during analyses. Deleteriousness/pathogenicity predictions tools for SNPs included Sorting Tolerant from Intolerant[132] (SIFT); the HumDIV algorithm from Polymorphism Phenotyping[133] (PolyPhen2); Combined Annotation Dependent Depletion[134] (CADD, version 1.3); and Functional Analysis through Hidden Markov Models[135] (FATHMM) using the FATHMM-MKL[136] algorithm. Annotation with ANNOVAR was implemented using a custom script outlined in Appendix 3. Pathogenicity and deleteriousness predictions for small insertions or deletions (indels) were manually

annotated with using CADD[134] and the FATHMM-indel[137] algorithm using these tools' online batch submission option.

2.2.4 Interrogating highly prioritised variants using the University of California Santa Cruz (UCSC) Genome Brower

To rank and further interrogate highly prioritised variants, the specific genomic positions were manually assessed on the University of California Santa Cruz (UCSC) Genome Brower (available at https://genome.ucsc.edu) to obtain evidence of functionality using the data from various tracks. The specific data and tracks included: DNaseI Hypersensitivity Clusters in 125 cell types from the ENCODE project[138] (V3), chromatin state segmentation by Hidden Markov Model (HMM) in nine human cell types from ENCODE/Broad,[138] and expression quantitative trait loci (eQTL) data from 44 tissues from Genotype-Tissue Expression (GTEx) Project[139, 140] (midpoint release, V6).

CHAPTER 3: SCREENING KERATOCONUS-CANIDATE GENES FOR RARE, PROTEIN-CODING POTENTIALLY PATHOGENIC VARIANTS IN A LARGE CASE-CONTROL COHORT

Publications arising from this chapter:

Lucas SEM, Zhou T, Blackburn NB, Mills RA, Ellis J, Leo P, Souzeau E, Ridge B, Charlesworth JC, Brown MA, Lindsay R, Craig JE, Burdon KP. **Rare, potentially pathogenic variants in 21 keratoconus candidate genes are not enriched in cases in a large Australian cohort of European descent**, *PLoS One*, 2018;13(6):e0199178.

Lucas SEM, Zhou T, Blackburn NB, Mills RA, Ellis J, Leo P, Souzeau E, Ridge B, Charlesworth JC, Brown MA, Lindsay R, Craig JE, Burdon KP. **Rare, potentially pathogenic variants in *ZNF469* are not enriched in keratoconus in a large Australian cohort of European descent**, *Investigative Ophthalmology and Visual Science* 2017;58(14):6248-56.

3.1 INTRODUCTION

To date, several approaches have been used to elucidate genetic variants that underpin keratoconus susceptibility and from this many candidate genes have been proposed to play a role in this disease. Linkage analysis in extended families has identified more than 20 linkage regions for keratoocus,[28, 29, 64-66, 69, 71, 72, 74-80, 141, 142] however, only regions on chromosome 5q have been replicated.[73-75, 141, 142] The number of loci identified highlights the heterogeneous nature of the disease. Such family-based studies have implicated few candidate variants and genes to date, due to the size of the regions. The most promising keratoconus gene identified with this method is *mir184*. This non-coding microRNA gene was found to have a pathogenic variant within the DNA binding domain in a family from Northern Ireland.[65, 77, 143] The same variant was subsequently identified in an unrelated family from Spain[69] and similar variants predicted to reduce the stability of the miRNA secondary structures were identified in two sporadic cases.[70] It is however important to note that these individuals had both keratoconus and congenital cataract and therefore may have a phenotype with a different genetic aetiology to isolated keratoconus. Variants in other genes, such as *IL1RN* and *SLC4A11*, have been hypothesised to play a role in disease due to the linkage-based identification of potentially pathogenic variants in an Ecuadorian family.[66] *IL1RN* which encodes IL1 receptor antagonist and *SLC4A11* which encodes solute carrier family 4 member 11 were selected as candidate genes due to their involvement in the immune response and apoptosis, respectively. However, these genes have not been assessed in other cohorts of keratoconus patients.

Genome-wide association studies (GWAS) have led to the identification of four genome-wide significant loci as well as several loci that show a suggestive association with keratoconus. A large GWAS that assessed loci associated with central corneal thickness (CCT) in keratoconus patients identified two single nucleotide polymorphisms (SNPs) associated with keratoconus in intronic regions of *FOXO1* (rs2721051) and *FND3CB* (rs4894535).[87] The same study found a suggestive association at rs1324183, between *MPDZ* and *NFIB*, which reached genome-wide significance following replication and meta-analysis.[87, 88] Similarly, rs4954218, upstream of *RAB3GAP1*, showed a suggestive association in the initial study, but reached significance after replication and meta-analysis.[85, 86] Other suggestive associations include SNPs in the promoter of *HGF*,[84, 144] rs1536482 between *RXRA* and *COL5A1*,[87] rs9938149 between *BANP* and *ZNF469*,[87] and two intronic SNPs (rs757219 and rs214884) in *IMMP2L*.[85] The identification of these loci has provided important insights into keratoconus genetics, however, functional variation at these loci have not yet been determined. While the most significant SNPs are located in non-coding regions, many of the nearby genes make good biological candidates for keratoconus. Thus, we hypothesise that rare protein-coding variation in these genes may be involved in keratoconus susceptibility.

The *BANP-ZNF469* locus has been highly controversial. The SNP, rs9938149, reached genome-wide significance with CCT and a suggestive association with keratoconus, however the genotype associated with a thinner cornea was associated with decreased keratoconus risk.[87] This finding was replicated in an independent cohort showing the same direction of association,[88] indicating that the association is likely to be real, if non-intuitive. As *ZNF469* is the closest gene to this SNP it has been hypothesised that genetic variation within the gene may account for the association at rs9938149 as well as contribute to CCT and keratoconus susceptibility. The potential role of *ZNF469* in keratoconus pathogenesis is further supported by the role of this gene in Brittle Cornea Syndrome type 1 (OMIM 229200). Brittle Cornea Syndrome type 1 is a rare, autosomal recessive connective tissue disorder, caused by biallelic loss-of-function variants in *ZNF469*. A key feature of this syndrome is extremely thin corneas that are prone to spontaneous rupture, suggesting that *ZNF469* is important for the structural integrity of the cornea. As an appealing candidate for keratoconus, coding variants in *ZNF469* have since been investigated. The findings of these studies are conflicting, with two of these studies reporting an association of potentially pathogenic variants in *ZNF469* with keratoconus,[145, 146] while two showed no association with disease.[147, 148]

Many genes have also been hypothesised to play a role in keratoconus based on their function and known corneal expression. The genes selected in the present study can broadly be categorised as regulatory genes, such as *CAST*[113] and *VSX1*;[89] structural genes, including the collagen genes *COL4A3* and *COL4A4*;[114] and genes involved in immune responses, such as *SOD1*,[103] *TF*[115] and *RAD51*,[116] *IL1A*[97] and *IL1B*.[117] The initial studies that implicated *CAST*, *COL4A3*, *COL4A1*, *TF*, *RAD51*, *IL1A* and *IL1B* in keratoconus showed associations at nearby or intronic SNPs. Therefore, the supporting evidence of

the involvement of these genes in disease has both a biological-, and positional-, basis. In contrast, the genes *VSX1* and *SOD1*, were initially proposed to play a role in keratoconus due to the identification of sequence variants in keratoconus patients that were absent in controls. As the first gene postulated to contribute to keratoconus, *VSX1* has been extensively assessed in many populations with conflicting results. Many studies conclude that *VSX1* is likely to be involved in keratoconus pathogenesis,[89-98, 149-152] while a similar number of studies do not find evidence of association.[99-102, 105, 153-161] Similarly, the superoxide dismutase gene (*SOD1*) has been screened in several populations including Slovenian,[102] Iranian,[95, 161] Italian,[96] Greek,[105] Saudi Arabian[106] and multiethnic[103] cohorts. A 7 bp intronic deletion was observed in cases but not controls in two of these reports[96, 103] and was significantly more frequent in cases compared to controls in another,[105] however, the remaining studies did not observe the variant.[95, 102, 106, 161] Given the contention surrounding the involvement of *VSX1* and *SOD1*, and the few studies that assessed the remaining functional candidates, further analysis is required to determine if they contribute to keratoconus susceptibility and pathogenicity.

Through these different methodologies and approaches, many candidate genes have been hypothesised to play a role in keratoconus based on their function, their proximity to associated SNPs or due to the identification of putatively causative variants within the gene. However, the majority of these genes have not been assessed beyond the initial study, and for those that have, the majority of studies have been small, with fewer than 100 keratoconus cases. To address this, our study assessed the role of 22 candidate genes in the largest cohort of keratoconus cases to date (n = 385), compared to 396 population controls. Specifically, our study examines the frequency of potentially pathogenic variants in *MPDZ*, *RXRA*, *RAB3GAP1*, *FOXO1*, *BANP*, *ZNF469*, *HGF*, *COL5A1*, *IMMP2L*, *FNDC3B*, *NFIB*, *ILRN*, *SLC4A11*, *CAST*, *COL4A3*, *COL4A4*, *TF*, *SOD1*, *VSX1*, *RAD51*, *IL1A* and *IL1B* in a cohort of Australians of European descent.

3.2 HYPOTHESIS AND AIM

The hypothesis that formed the foundation of this study was that rare, highly penetrant protein-coding variants contribute to keratoconus development. More specifically, it was hypothesised that candidate genes for keratoconus are enriched in rare protein-coding variants, predicted to be pathogenic (or similar) by *in silico* tools, in cases compared to controls. This led to the following aim:

To determine if potentially pathogenic variants are enriched in known candidate genes in keratoconus patients, compared to controls.

3.3 METHODS

3.3.1 Study participants

The case cohort consisted of 385 keratoconus patients of European descent (detailed in Section 2.1.1). The control cohort consisted of 396 ethnically matched females from the Anglo-Australasian Osteoporosis Genetics Consortium (AOGC; described in Section 2.1.2). These control individuals were not assessed for eye disease, and therefore are expected to have the population frequency of keratoconus. For this reason, this cohort is described as 'population controls'.

Due to coverage issues across *ZNF469* in the population control cohort, a second control cohort of 230 Australians of European descent was used for the analysis of this candidate gene. These individuals were screened for keratoconus and were found to be unaffected but were affected by other eye diseases (largely glaucoma). A full description of this cohort is provided in Section 2.1.3. Throughout this chapter, this cohort is referred to as the 'screened controls'.

3.3.2 Whole exome sequencing data

Whole exome sequencing (WES) data were available for 99 keratoconus cases. WES was conducted by Macrogen Inc. using the SureSelect Human All Exon V4 enrichment kits (Agilent Technologies Inc., Santa Clara, California, USA) with paired-end sequencing on an Illumina HiSeq 2000 (Illumina, San Diego, California, USA). The Churchill pipeline[162] was used to align raw reads to hg19 using BWA-MEM[163] (version 0.7.12) and variants were joint-called with SAMtools[164] (version 1.3.1) and BCFtools (version 1.3.1; (https://github.com/samtools/BCFtools). The 230 individuals in the screened control cohort (ANZRAG) were sequenced and analysed in parallel with the keratoconus cases, however, these data were only used for analysis of *ZNF469*.

WES was generated for the population control cohort (AOGC) using Illumina's TruSeq Exome Enrichment on an Illumina HiSeq 2000 at the University of Queensland Centre for Clinical Genomics. Raw reads were aligned to the human reference genome (hg19) using novoalign (version 2.08; http://www.novocraft.com/products/novoalign/) and variant calling and quality score calibration was conducted using GATK[125] (version 3.2-2), according to GATK's 'Best Practices Guidelines'.[165, 166] These data were provided by Professor Matthew Brown (Queensland University of Technology and Translational Research Institute, Princess Alexandra Hospital, QLD, Australia).

3.3.3 Targeted gene screen in additional cases

The protein-coding regions of 22 genes of interest were sequenced using a targeted sequencing approach in a total of 341 cases across 44 DNA pools using the HaloPlex Target Enrichment System (Agilent Technologies). The selected genes are presented in Table 3.1. A custom probe panel was designed using Agilent Technologies' SureDesign Custom Design Tool. DNA samples were quantitated using double

stranded DNA (dsDNA) quantitation assays on either a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) using a Qubit dsDNA High Sensitivity Assay Kit or on a Fluoroskan plate reader (Thermo Fisher Scientific) with a Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, Carlsbad, California, USA). DNA pools containing equimolar DNA samples from eight keratoconus patients were prepared as published previously.[84] Using Agilent's HaploPlex Target Enrichment System Protocol for Illumina Sequencing (version D.5, May 2013), each DNA pool was indexed with a unique indexing primer cassette, allowing for multiplexed sequencing. Enriched libraries were validated and quantified using Agilent's Bioanalyzer High Sensitivity DNA Assay Kit on a 2100 Bioanalyzer (Agilent Technologies Inc.). Sequencing was conducted in batches of 11 DNA pools on the MiSeq platform (Illumina) using a MiSeq V2 Reagent kit (300 cycles) with paired-end reads. SureCall (Agilent Technologies Inc.) was used for sequencing analysis using standard trimmer parameters, the BWA-MEM algorithm to align reads and the SNPPET SNP Caller (part of SureCall) to call variants. Variants were called if a minimum read depth of 10 and quality score of 20 was reached. As variants were called from pooled DNA samples it was expected that if a single alternate allele was present it would be observed on 6.25% of the reads mapping to that position. To account for this, the minimum allele frequency for heterozygous variants was set to 0.035 and the threshold for calling insertions and deletion variants (indels) was ≥ 0.04. It is important to note that 55 cases included in this pooled gene screen were included in the whole exome sequencing; providing additional cross-validation between the two sequencing strategies, however, variants from these samples were only counted in the analyses once.

Table 3.1 – Genes included in the targeted gene screen.

Gene	Transcript
COL4A3	NM_000091
COL4A4	NM_000092
IL1A	NM_000575
IL1B	NM_000576
IL1RN	NM_173841
RAB3GAP1	NM_012233
TF	NM_001063
FNDC3B	NM_022763
CAST	NM_00104244
HGF	NM_000601
IMMP2L	NM_032549
COL5A1	NM_000093
NFIB	NM_005596
MPDZ	NM_003829
RXRA	NM_00129192
FOXO1	NM_002015
RAD51	NM_002875
BANP	NM_017869
ZNF469	NM_00112746
SLC4A11	NM_032034
VSX1	NM_014588
SOD1	NM_000454

3.3.4 Included genomic regions and variant annotation

To ensure a sound comparison when comparing the frequency of variants between data with different capture methods, only target regions that were common to all three capture methods (exome captures for the cases and population controls, as well as, the targeted sequencing) with a mean read depth ≥10 were included in the analysis. For the WES data, all included individuals had high confidence genotypes for ≥ 90% of the included regions. Due to limited coverage of *ZNF469* in the population control cohort and the recent interest in this gene in keratoconus susceptibility, *ZNF469* was investigated separately as outlined in Section 3.3.10.

Variants identified within the included regions were annotated with ANNOVAR[127] as described in detail in Section 2.2.3. Notably, the variants were annotated with the minor allele frequency (MAF) observed in the non-Finnish European population of the Exome Aggregation Consortium database[129]

(ExAC NFE) and pathogenicity/deleteriousness predictions from Sorting Tolerant from Intolerant[132] (SIFT), the HumDIV algorithm from Polymorphism Phenotyping[133] v2 (PolyPhen2) and Combined Annotation–Dependent Depletion[134] (CADD; version 1.3).

3.3.5 Filtering strategy to identify potentially pathogenic variants

Variants were only included in analyses if a sequencing depth of ≥10 and a quality score of ≥20 was obtained. Variants meeting these thresholds were then filtered to include single nucleotide polymorphsism (SNPs) that were predicted to be pathogenic by SIFT or PolyPhen2 with a MAF <0.01. In addition, SNPs with a scaled CADD score ≥15 that met the MAF threshold were included in this filtering strategy. These variants were considered 'potentially pathogenic variants' and were included in our statistical analysis. Insertions and deletions were not included.

3.3.6 Determining thresholds for variant inclusion from the pooled gene screen

Extensive assessment of variants in the pooled gene screen dataset was conducted to determine thresholds to classify variants called as real or artefact, as well as to determine the likely number of alternate alleles present in the DNA pool. The information in the VCF file, including the frequency of a variant call across DNA pools, along with manual inspection of the reads mapping to a variant in the SureCall Triage View, were used to predict whether unassessed variants were likely to be real or sequencing artefacts. Suspected artefacts and real variants were selected for validation by direct sequencing in all individuals included in the DNA pool (described in Section 3.3.7), to identify thresholds for the inclusion of variants.

3.3.7 Variant validation by direct sequencing

To validate variants identified in the case cohort, primers were designed using Primer3Plus[167] or PrimerBlast.[168] DNA was amplified with MyTaq HS Mix (Bioline, London, UK) and purified using either Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA) according to the manufacturer's instructions, or equivalent magnetic beads prepared in-house. Purified amplicons were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) on an ABI 310 or ABI 3500 (Applied Biosystems). DNA sequences were aligned to the human reference genome (hg19), and chromatograms were manually inspected at the position of each variant using Sequencher 4.10.1 (http://www.genecodes.com/; Gene Codes Corporation, Ann Arbor, MI, USA) to assess validation.

3.3.8 Comparing variant calls between WES and the pooled gene screen data

Fifty-five individuals were included in both the WES and pooled gene screen datasets. Potentially pathogenic variants called in these individuals were compared between these two methods to assess the

consistency of variants calls and the utility of the pooled targeted gene screen. Variants were only included in the statistical analysis once.

3.3.9 Statistical analysis

Following variant filtering, genes with at least one alternate allele identified in the case cohort were included in the statistical analysis. For each of these genes separately, the number of potentially pathogenic variants were compared between cases and controls using a Yates corrected chi squared test or a Fishers' exact test, where appropriate. Odds ratios and 95% confidence intervals were also calculated. As 21 genes (not including *ZNF469*, see Section 3.3.10) were assessed in the present study, a significance threshold of $p < 0.0024$ (0.05/21) was determined using the Bonferroni correction for multiple testing.

3.3.10 Additional investigations of rare protein-coding variants in *ZNF469*

A detailed analysis of potentially pathogenic variants in *ZNF469* was conducted separately due to limited coverage of the gene in the population controls, as well as, previous studies of the gene providing a strong *a priori* hypothesis for the association of rare protein-coding variants in *ZNF469* with keratoconus.

3.3.10.1 Genomic regions included in the analysis of *ZNF469*

For the analysis for *ZNF469*, potentially pathogenic variants were compared between cases and two control cohorts (the population controls and the screened controls) in separate comparisons. Only regions with sufficient coverage across the relevant datasets were included in each comparison to ensure high quality sequence data and robust analysis. While it has recently been suggested that *ZNF469* is a single-exon gene,[169] the capture methods in this study were designed under the assumption that *ZNF469* has two exons with a short 84 bp intron. Therefore, only exonic variants defined by transcript NM_001127464 were included in the analysis. In contrast with the other 21 genes, this analysis did not exclude indels from the statistical analysis. Annotation of variants identified in *ZNF469* was performed as described in Section 2.2.3.

3.3.10.2 Variant filtering strategies for *ZNF469*

Two filtering strategies were used to define rare and very rare potentially pathogenic variants to make the results comparable to previous studies. Filtering Strategy 1 included nonsynonymous variants with a MAF < 0.01 (rare) with a damaging or probably/possibly damaging prediction by SIFT and/or PolyPhen2. As SIFT and PolyPhen2 do not assess indels, all exonic indels with a MAF < 1% were included in Filtering Strategy 1. Filtering Strategy 2 was identical to the first filtering strategy, except variants were required to have a MAF < 0.001 (very rare). These filtering strategies were designed to

allow direct comparison with previous studies and therefore did not include CADD predictions of deleteriousness like the other 21 genes included in this study.

3.3.10.3 Statistical analyses for *ZNF469*

For both filtering strategies, the number of alternate alleles was compared between cases and the two control datasets separately using chi squared or Fisher's exact tests, where appropriate. The odds ratio (OR) and 95% confidence interval (95% CI) were calculated. In addition, the Sequence Kernel Association Test (SKAT)[170] was used to assess if potentially pathogenic variants were enriched in the case WES data compared to the two control cohorts. For this analysis, all nonsynonymous variants predicated to be potentially pathogenic, regardless of MAF, were included. Using the two control cohorts as a separate comparison, SKATBinary was run using the quantile adjusted moment matching (QA) method and the default weight parameter, Beta (1, 25). This weighting applies strong weights to rare variants, non-zero weighting to uncommon variants (MAF 0.01 – 0.05) and almost zero weights to common variants.[170]

3.3.10.4 Genetic power calculations for *ZNF469*

Power calculations were conducted using the 'case-control for discrete traits' module of the genetic power calculator.[171] These calculations assumed an additive model where the prevalence of keratoconus was 0.00067 (1 in 1500) and D' prime was equal to 1. The high-risk allele frequency was set for each filtering strategy separately, using the frequency of variants identified in the screened controls.

3.3.10.5 Data visualisation for *ZNF469*

Coverage across *ZNF469* was plotted for the WES datasets based on the mean depth at variant positions using the R[126] package ggplot2.[172] To demonstrate the differences in coverage across the entire gene, this plot was aligned to a schematic of *ZNF469* outlining the position of the exons and zinc finger motifs (based on the positions obtained from UniProt;[173] entry Q96JG9) using the R package cowplot.[174]

Variants included in Filtering Strategy 1 were plotted as a bar plot for each group using custom R scripts and the ggplot2 package. Each variant was plotted along the x-axis according to its genomic position and the frequency of the variant in the study group was indicated by the height of the bar. Indels were plotted according to the position of the first affected base. To indicate the conservation of each variant position, the bars were coloured using a colour gradient corresponding to the 100-way vertebrate PhastCons[175, 176] score as available from the University of Santa Cruz (UCSC) Genome Browser.[177] Similarly to the coverage plot described above, the plots for each group were aligned to a schematic of *ZNF469* using the R package cowplot, to allow for comparison of the variants across the gene.

3.4 RESULTS

3.4.1 Determining thresholds for variant inclusion from the pooled gene screen

A total of 46 variants called in the pooled gene screen, including 27 variants in *ZNF469*, were assessed for validation to identify thresholds for the inclusion of variants or the number of alternate alleles present in the DNA pool. Evidence that a variant was an artefact included: the variant was present in three or more pools, despite having a MAF < 0.001 or no dbSNP ID; the variant was observed on an unexpectedly high frequency of the reads in a single pool (ie. > 40%); the variant was only present in a single population of amplicons; the variant was called in a region with low coverage (ie < 20 reads); the variant was present on reads with multiple spurious variants; or the variant had previously been sequenced by direct sequencing in another DNA pool and was identified as an artefact. If, after collating the evidence, it was still not clear if the variant was likely to be real or not, the variant was validated by direct sequencing. Based on the selected validation experiment, variants deemed to be real with a DNA pool frequency of 3.7% – 10.6% were predicted to have a single alternate allele, whereas two alternate alleles were predicted for variants with a frequency between 10.7% and 15.5%.

3.4.2 Comparing variant calls between WES and the pooled gene screen data

Across all 22 genes of interest, 34 variants were identified 35 times in the WES data, including two small deletions and 12 SNPs identified in *ZNF469*. Thirty of these variants were also were identified in the pooled gene screen data, with four SNPs and one small deletion (in *ZNF469*) not meeting the minimum alternate allele threshold of 3.5% of reads. This alternate allele threshold of 3.5% was, however, carefully selected to minimise the inclusion of spurious variant calls. In contrast, a *ZNF469* variant was identified in an individual that was included in both sequencing strategies through the targeted gene screen that was not called in the WES dataset. Upon manual investigation of the WES data, the variant was in fact called, but had a depth of only five reads and was thus not included in the filtered data. This variant was therefore included in the chi squared test which included the pooled gene screen data but was excluded in the SKAT analysis which only used the WES data.

3.4.3 Rare potentially pathogenic variants in 21 candidate genes

Demographic details for keratoconus cases and controls are provided in Table 3.2.

Table 3.2 – Demographics of keratoconus cases and controls at the time of examination.

Cohort	n	Mean age* (range)	% Female	Disease status
Cases	385	45.2 (14-85)	44.2	affected
Controls	396	69.7 (46-86)	100	unscreened

n = the number of individuals
* Age is reported in years

Coverage statistics for each of the 21 genes included in the analysis are summarised for the protein-coding portions of the longest transcript that best fit the capture designs in Table 3.3. A full list of the included regions for these gene are presented in the supporting information of the published article (available at: https://doi.org/10.1371/journal.pone.0199178.s001). Due to a high GC content, the coverage of *BANP* was poor, particularly in the control data and the targeted sequencing dataset. When considering regions captured by probes across the three datasets, only 58% of the gene was captured, however, 70% of these regions had sufficient coverage for analysis. Similarly, 77.4% of *VSX1* was captured across all datasets due to insufficient capture of GC-rich regions and 60.6% of these captured regions met the coverage threshold for inclusion in the analysis. Despite this poor coverage, a number of previously reported *VSX1* variants in keratoconus were sufficiently covered for analysis. The captured regions of the remaining genes ranged from 75.3 – 97.5% and, apart from *FOXO1*, 99% - 100% of the coding bases in captured regions were included in analysis. For *FOXO1*, the included portion of the captured region dropped to 82.5% as part of the first exon did not meet the minimum depth threshold for variant calling in the WES datasets. It is important to note that additional regions, including non-protein-coding regions, may also have been included in analysis.

Table 3.3 – Coverage statistics for each gene of interest.

Gene	Coding bases	Captured coding bases (%)	Analysed coding bases (%)
COL4A3	5,013	4,136 (82.5)	4,096 (99.0)
COL4A4	5,073	4,200 (82.8)	4,184 (99.6)
IL1A	816	762 (93.4)	762 (100)
IL1B	810	629 (77.7)	629 (100)
IL1RN	543	499 (91.9)	499 (100)
RAB3GAP1	2,946	2,797 (94.9)	2,797 (100)
TF	2,233	1,683 (75.3)	1,681 (99.9)
FNDC3B	3,615	3,015 (83.4)	3,015 (100)
CAST	2,310	1,978 (85.6)	1,968 (99.5)
HGF	2,187	2,116 (96.8)	2,116 (100)
IMMP2L	528	501 (94.9)	501 (100)
COL5A1	5,517	4,550 (82.5)	4,548 (100)
NFIB	1,263	1,068 (84.6)	1,068 (100)
MPDZ	6,126	5,534 (90.3)	5,477 (99.0)
RXRA	1,098	940 (85.6)	940 (100)
FOXO1	1,968	1,893 (96.2)	1,561 (82.5)
RAD51	1,020	928 (91.0)	928 (100)
BANP	1,410	818 (58.0)	573 (70.0)
SLC4A11	2,676	2,610 (97.5)	2,610 (100)
VSX1	1,098	850 (77.4)	515 (60.6)
SOD1	465	404 (86.9)	404 (100)

Coding bases = the total number of bases in the protein-coding portion of the transcript.
Captured coding bases = the total number of the protein-coding bases included in the captured regions (and percentage of the total number of protein-coding bases in the transcript).
Analysed coding bases = the total number of protein-coding bases included in the analysis (and the percentage this represents of the captured protein-coding bases).

Following variant filtering, 164 potentially pathogenic variants were identified across both cases and controls (Table 3.4). This included 138 nonsynonymous variants, 21 synonymous variants, three nonsense variants, one intronic variant and one variant in a 3' untranslated region. Of these variants, 70 were unique to the cases, 69 were only observed in the population controls and 25 were identified in both groups. For the cases, a total of 146 potentially pathogenic variants were observed, while 192 were identified in population controls. Two variants, p.(R85Q) in *IL1A* and p.(P1379S) in *COL5A1*, were each observed in the homozygous state in a single control. All other variants were observed as heterozygotes.

No potentially pathogenic variants were identified in *IL1RN* in either cases or controls. Additionally, potentially pathogenic variants were not observed in the case cohort in *BANP*, *IL1B*, *RAD51* or *SOD1*. However, in the control group, one potentially pathogenic variant was observed in both *IL1B* and *SOD1*, two variants were identified in *BANP*, and three were observed in *RAD51*. For the remaining genes, the total number of potentially pathogenic variants identified across both groups ranged from three in *RXRA* up to 102 in *MPDZ*. Of the genes included in the chi-square or Fishers' exact tests, *COL4A3* and *MPDZ* showed a nominally higher frequency of potentially pathogenic variants in controls compared to the case cohort, with both genes obtaining a $p < 0.05$ in the burden analysis, however, neither gene remained significant under correction for multiple testing. All other genes showed no difference between groups. Statistics for each gene included in the analysis are summarised in Table 3.5.

Table 3.4 – Potentially pathogenic variants identified across the 21 genes of interest.

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
IL1A	chr2:113539246	c.254G>A	p.(R85Q)	rs3783531	D (0.045)	D (1.000)	25.00	0.0024	0.0152 (6)	0.0052 (2)
IL1A	chr2:113540315	c.74T>C	p.(I25T)	rs139798825	D (0.004)	P (0.500)	20.60	0.0008	0.0051 (2)	0.0026 (1)
IL1B	chr2:113588006	c.742G>A	p.(V248I)	rs781114719	D (0.026)	D (0.991)	28.10	<0.0001	0.0025 (1)	0.0000 (0)
RAB3GAP1	chr2:135878409	c.669G>T	p.(L223F)	rs76927619	D (0.013)	D (0.993)	23.30	<0.0001	0.0000 (0)	0.0026 (1)
RAB3GAP1	chr2:135883787	c.867C>T	p.(T289=)	NA	NA	NA	16.20	0.0000	0.0000 (0)	0.0026 (1)
RAB3GAP1	chr2:135887597	c.1006C>T	p.(R336C)	rs150478342	T (0.148)	B (0.014)	24.30	0.0067	0.0101 (4)	0.0052 (2)
RAB3GAP1	chr2:135911271	c.2114G>A	p.(R705Q)	rs367558491	D (0.006)	D (1.000)	34.00	<0.0001	0.0000 (0)	0.0026 (1)
RAB3GAP1	chr2:135926239	c.2834T>G	p.(V945G)	NA	D (0.017)	D (0.998)	27.20	0.0000	0.0025 (1)	0.0000 (0)
COL4A4	chr2:227872132	c.4982T>A	p.(F1661Y)	rs374119389	D (0.008)	P (0.830)	23.30	0.0003	0.0025 (1)	0.0000 (0)
COL4A4	chr2:227872153	c.4961C>T	p.(T1654M)	rs771066050	D (0.000)	D (1.000)	23.20	<0.0001	0.0025 (1)	0.0000 (0)
COL4A4	chr2:227872783	c.4760C>G	p.(P1587R)	rs190148408	T (0.902)	P (0.913)	0.90	0.0027	0.0025 (1)	0.0052 (2)
COL4A4	chr2:227872812	c.4731G>A	p.(A1577=)	rs200639109	NA	NA	17.00	0.002	0.0000 (0)	0.0026 (1)
COL4A4	chr2:227872894	c.4649C>G	p.(P1550R)	NA	D (0.000)	D (0.998)	22.60	0.0000	0.0025 (1)	0.0000 (0)
COL4A4	chr2:227875029	c.4522G>A	p.(G1508S)	NA	D (0.000)	D (1.000)	23.70	0.0000	0.0025 (1)	0.0000 (0)
COL4A4	chr2:227875065	c.4486C>A	p.(L1496M)	NA	T (0.069)	P (0.760)	12.30	0.0000	0.0000 (0)	0.0026 (1)
COL4A4	chr2:227896939	c.3631G>A	p.(E1211K)	rs750501128	D (0.03)	B (0.006)	23.30	<0.0001	0.0000 (0)	0.0026 (1)
COL4A4	chr2:227912248	c.3232G>A	p.(A1078T)	rs77277077	T (0.422)	D (0.988)	23.60	0.0001	0.0025 (1)	0.0000 (0)
COL4A4	chr2:227924228	c.2276C>T	p.(P759L)	rs36121515	D (0.014)	P (0.911)	25.40	0.0004	0.0051 (2)	0.0000 (0)
COL4A4	chr2:227945181	c.1781A>G	p.(E594G)	rs35998949	T (0.434)	B (0.003)	20.60	0.0003	0.0000 (0)	0.0026 (1)
COL4A4	chr2:227964372	c.1063C>G	p.(P355A)	rs368293426	T (0.308)	D (1.000)	26.20	<0.0001	0.0000 (0)	0.0026 (1)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
COL4A4	chr2:227973309	c.723A>C	p.(Q241H)	rs201673987	T (0.056)	D (0.990)	20.40	<0.0001	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228110691	c.346C>A	p.(P116T)	rs115324397	D (0.027)	D (1.000)	24.70	0.0082	0.0429	0.0260
COL4A3	chr2:228110718	c.373T>C	p.(C125R)	NA	T (0.051)	D (0.999)	27.80	0.0000	0.0000 (0)	0.0026 (1)
COL4A3	chr2:228113204	c.514G>A	p.(D172N)	rs377575924	T (0.098)	D (0.996)	34.00	<0.0001	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228118867	c.805G>A	p.(E269K)	rs80109666	T (0.921)	P (0.827)	0.01	0.0046	0.0000 (0)	0.0052 (2)
COL4A3	chr2:228124533	c.1054G>T	p.(E352*)	NA	NA	NA	36.00	0.0000	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228131783	c.1483C>T	p.(H495Y)	rs200510532	D (0.040)	D (0.985)	22.30	0.0013	0.0025 (1)	0.0026 (1)
COL4A3	chr2:228141146	c.1973C>T	p.(P658L)	rs770397467	D (0.002)	D (0.983)	23.50	<0.0001	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228153949	c.2965C>T	p.(P989S)	rs774477588	T (0.648)	P (0.896)	21.60	<0.0001	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228163401	c.3755C>T	p.(A1252V)	rs761179248	T (0.346)	B (0.000)	19.80	<0.0001	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228163475	c.3829G>A	p.(G1277S)	rs190598500	D (0.000)	D (1.000)	23.50	0.0005	0.0000 (0)	0.0026 (1)
COL4A3	chr2:228172594	c.4421T>C	p.(L1474P)	rs200302125	D (0.000)	D (1.000)	23.40	0.0046	0.0227 (9)	0.0026 (1)
COL4A3	chr2:228173618	c.4466C>T	p.(T1489I)	rs200818438	T (0.140)	D (0.997)	25.50	<0.0001	0.0025 (1)	0.0000 (0)
COL4A3	chr2:228175629	c.4893C>T	p.(F1631=)	rs183218622	NA	NA	16.60	0.0053	0.0152 (6)	0.0130 (5)
TF	chr3:133474263	c.559C>G	p.(P187A)	rs751656601	T (0.100)	B (0.101)	23.60	<0.0001	0.0000 (0)	0.0026 (1)
TF	chr3:133475754	c.771C>A	p.(Y257*)	NA	NA	NA	26.30	0.0000	0.0000 (0)	0.0026 (1)
TF	chr3:133485143	c.1352T>C	p.(V451A)	rs142116896	T (0.197)	P (0.582)	20.40	0.0001	0.0025 (1)	0.0000 (0)
TF	chr3:133485196	c.1405T>C	p.(C469R)	NA	D (0.000)	D (1.000)	25.80	0.0000	0.0000 (0)	0.0026 (1)
TF	chr3:133489405	c.1676A>T	p.(Q559L)	rs753682414	D (0.000)	P (0.793)	23.90	0.0001	0.0000 (0)	0.0026 (1)
TF	chr3:133496032	c.2012G>A	p.(G671E)	rs121918677	D (0.000)	D (1.000)	28.90	0.0037	0.0202 (8)	0.0104 (4)
FNDC3B	chr3:171830293	c.24C>G	p.(T8=)	rs539042799	NA	NA	16.20	0.0000	0.0000 (0)	0.0026 (1)
FNDC3B	chr3:171965444	c.386A>G	p.(H129R)	NA	T (0.822)	D (0.989)	22.90	0.0000	0.0000 (0)	0.0026 (1)
FNDC3B	chr3:172046816	c.1329G>A	p.(P443=)	rs372880783	NA	NA	15.20	<0.0001	0.0025 (1)	0.0000 (0)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
FNDC3B	chr3:172052819	c.1727C>T	p.(T576I)	rs757447664	T (0.369)	P (0.818)	23.40	<0.0001	0.0000 (0)	0.0026 (1)
FNDC3B	chr3:172070668	c.2590G>A	p.(V864I)	rs143064249	T (0.098)	B (0.063)	18.00	0.0023	0.0025 (1)	0.0026 (1)
FNDC3B	chr3:172096114	c.3063G>A	p.(T1021=)	rs561549306	NA	NA	15.20	<0.0001	0.0000 (0)	0.0026 (1)
CAST	chr5:96031558	c.157G>A	p.(G53S)	rs199633496	D (0.048)	P (0.559)	8.70	0.0002	0.0000 (0)	0.0026 (1)
CAST	chr5:96031567	c.166C>A	p.(Q56K)	rs780010534	D (0.028)	D (0.982)	23.00	0.0000	0.0000 (0)	0.0026 (1)
CAST	chr5:96031601	c.200C>T	p.(S67L)	rs776197495	T (0.274)	P (0.625)	12.30	0.0000	0.0025 (1)	0.0000 (0)
CAST	chr5:96073629	c.710C>T	p.(T237M)	rs779432064	T (0.107)	D (0.994)	13.40	<0.0001	0.0025 (1)	0.0000 (0)
CAST	chr5:96078437	c.1111C>T	p.(R371C)	rs377565522	D (0.044)	P (0.879)	23.50	<0.0001	0.0000 (0)	0.0026 (1)
CAST	chr5:96078456	c.1130A>G	p.(D377G)	rs144157006	T (0.066)	D (0.999)	23.60	0.0009	0.0051 (2)	0.0000 (0)
CAST	chr5:96079350	c.1134+890C>T	NA	rs780684334	D (0.004)	D (0.984)	17.00	<0.0001	0.0025 (1)	0.0000 (0)
CAST	chr5:96082128	c.1217C>T	p.(T406M)	rs78054235	T (0.201)	D (0.986)	12.60	0.0001	0.0025 (1)	0.0000 (0)
CAST	chr5:96089833	c.1528G>C	p.(E510Q)	rs773300671	D (0.046)	D (0.997)	26.70	0.0000	0.0000 (0)	0.0026 (1)
CAST	chr5:96090383	c.1582C>T	p.(R528C)	rs142879548	T (0.073)	D (1.000)	26.10	0.0007	0.0025 (1)	0.0000 (0)
CAST	chr5:96090425	c.1624T>C	p.(Y542H)	rs140928951	D (0.003)	D (1.000)	25.40	<0.0001	0.0000 (0)	0.0026 (1)
CAST	chr5:96100929	c.1871G>A	p.(S624N)	rs149919102	T (0.982)	P (0.741)	7.80	0.0000	0.0000 (0)	0.0026 (1)
HGF	chr7:81335011	c.1816A>G	p.(I606V)	rs765890500	T (1.000)	P (0.814)	10.40	<0.0001	0.0025 (1)	0.0000 (0)
HGF	chr7:81335013	c.1814C>T	p.(T605I)	rs147075806	T (0.520)	B (0.000)	16.50	<0.0001	0.0025 (1)	0.0000 (0)
HGF	chr7:81346592	c.1361C>A	p.(T454K)	NA	D (0.000)	D (1.000)	31.00	0.0000	0.0000 (0)	0.0026 (1)
HGF	chr7:81358978	c.983G>A	p.(R328H)	rs374484762	D (0.030)	B (0.296)	28.30	<0.0001	0.0000 (0)	0.0026 (1)
HGF	chr7:81392140	c.137C>T	p.(A46V)	rs150267054	T (0.229)	D (0.998)	24.40	0.003	0.0076 (3)	0.0026 (1)
HGF	chr7:81399282	c.6G>T	p.(W2C)	rs745851853	D (0.002)	D (0.999)	34.00	0.0000	0.0025 (1)	0.0000 (0)
IMMP2L	chr7:111161438	c.66G>A	p.(A22=)	rs148496605	NA	NA	20.20	0.003	0.0076 (3)	0.0000 (0)
IMMP2L	chr7:111161439	c.65C>T	p.(A22V)	rs202048991	T (0.247)	D (0.978)	23.50	0.0003	0.0000 (0)	0.0026 (1)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
MPDZ	chr9:13107088	c.6002G>A	p.(R2001H)	rs1802495	T (0.129)	D (1.000)	34.00	0.0002	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13108980	c.5934C>G	p.(S1978R)	rs758152471	D (0.034)	D (1.000)	28.30	0.0007	0.0076 (3)	0.0000 (0)
MPDZ	chr9:13109969	c.5837T>C	p.(I1946T)	rs201230061	T (0.508)	D (0.984)	26.40	0.0001	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13119542	c.5338G>A	p.(V1780I)	rs202112833	T (0.251)	D (1.000)	31.00	0.0002	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13122091	c.5032T>G	p.(L1678V)	rs763372118	T (0.06)	P (0.746)	25.60	<0.0001	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13126531	c.4616G>A	p.(G1539D)	rs779148372	D (0.018)	D (1.000)	34.00	0.0002	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13126771	c.4465G>A	p.(D1489N)	NA	D (0.017)	D (0.999)	27.20	0.0000	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13136096	c.4378A>G	p.(K1460E)	rs767920289	T (0.313)	B (0.174)	17.30	<0.0001	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13136715	c.4288A>G	p.(I1430V)	rs765489877	T (0.541)	B (0.078)	16.30	0.0001	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13136783	c.4220G>A	p.(Y1407C)	rs200891478	D (0.049)	D (1.000)	28.90	0.0091	0.0101 (4)	0.0026 (1)
MPDZ	chr9:13138020	c.4136G>A	p.(G1379E)	NA	T (0.159)	D (1.000)	33.00	0.0000	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13139994	c.3995A>G	p.(Y1332C)	NA	T (0.105)	D (0.991)	25.30	0.0000	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13140070	c.3919G>A	p.(E1307K)	rs199509495	T (0.886)	D (0.967)	18.30	0.0013	0.0101 (4)	0.0026 (1)
MPDZ	chr9:13143469	c.3836A>G	p.(D1279G)	rs370624740	T (0.146)	B (0.043)	16.70	<0.0001	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13147602	c.3686G>A	p.(R1229Q)	rs61753783	T (0.106)	D (1.000)	35.00	0.0005	0.0051 (2)	0.0000 (0)
MPDZ	chr9:13147629	c.3659G>A	p.(S1220N)	NA	T (0.086)	D (0.998)	28.60	0.0000	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13150558	c.3582T>G	p.(S1194R)	rs188840960	D (0.033)	D (1.000)	25.70	0.0079	0.0227 (9)	0.0052 (2)
MPDZ	chr9:13150617	c.3523C>T	p.(R1175W)	rs367915328	D (0.004)	D (1.000)	29.70	<0.0001	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13158017	c.3452G>A	p.(R1151Q)	rs367828845	T (0.356)	D (0.999)	28.90	<0.0001	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13158026	c.3443A>G	p.(Q1148R)	rs752969006	T (0.238)	D (0.982)	23.20	0.0000	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13158062	c.3407G>A	p.(S1136N)	rs41265286	T (0.144)	D (1.000)	30.00	0.0028	0.0051 (2)	0.0000 (0)
MPDZ	chr9:13162786	c.3263A>G	p.(Y1088C)	rs372125677	D (0.009)	P (0.900)	24.30	<0.0001	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13168369	c.3250A>C	p.(I1084L)	NA	T (0.435)	B (0.384)	22.20	0.0000	0.0000 (0)	0.0026 (1)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
MPDZ	chr9:13168435	c.3184A>G	p.(T1062A)	rs763919393	T (0.291)	D (0.983)	24.20	<0.0001	0.0000 (0)	0.0026 (1)
MPDZ	chr9:13175852	c.2954A>C	p.(Q985P)	rs200272559	D (0.026)	B (0.371)	12.30	0.0001	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13183529	c.2537C>G	p.(S846C)	rs200553028	D (0.012)	D (1.000)	25.00	0.0011	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13186355	c.2395A>G	p.(K799E)	rs150038177	T (0.084)	P (0.911)	21.50	0.0067	0.0025 (1)	0.0130 (5)
MPDZ	chr9:13188921	c.2226T>A	p.(D742E)	NA	D (0.029)	D (1.000)	24.80	0.0000	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13188953	c.2194T>A	p.(S732T)	rs200475640	T (0.085)	D (0.999)	24.60	0.0021	0.0076 (3)	0.0026 (1)
MPDZ	chr9:13190162	c.2105A>T	p.(E702V)	rs4740548	D (0.004)	D (0.994)	28.30	0.009	0.0227 (9)	0.0208 (8)
MPDZ	chr9:13190163	c.2104G>A	p.(E702K)	rs4741289	D (0.005)	D (0.990)	33.00	0.009	0.0227 (9)	0.0208 (8)
MPDZ	chr9:13190165	c.2102T>C	p.(I701T)	NA	T (0.085)	B (0.372)	23.80	0.0000	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13219751	c.893G>C	p.(S298T)	rs201889514	T (1.000)	D (1.000)	25.50	0.0005	0.0025 (1)	0.0026 (1)
MPDZ	chr9:13221466	c.781G>C	p.(V261L)	rs200611423	T (0.071)	B (0.270)	23.00	0.0002	0.0025 (1)	0.0000 (0)
MPDZ	chr9:13223592	c.511C>G	p.(Q171E)	rs181479224	D (0.030)	D (0.995)	23.80	0.0029	0.0076 (3)	0.0000 (0)
MPDZ	chr9:13224441	c.325G>T	p.(G109C)	rs61753782	T (0.132)	D (0.990)	5.50	0.0074	0.0000 (0)	0.0052 (2)
NFIB	chr9:14088119T	c.*189A>T	NA	rs548618850	NA	NA	18.20	0.0000	0.0025 (1)	0.0000 (0)
NFIB	chr9:14125753	c.938C>T	p.(P313L)	NA	D (0.001)	D (1.000)	34.00	0.0000	0.0025 (1)	0.0000 (0)
NFIB	chr9:14150247	c.703A>G	p.(T235A)	NA	D (0.041)	B (0.000)	15.10	0.0000	0.0000 (0)	0.0026 (1)
NFIB	chr9:14307342	c.208C>G	p.(Q70E)	NA	D (0.020)	P (0.843)	24.30	0.0000	0.0025 (1)	0.0000 (0)
NFIB	chr9:14307354	c.196C>T	p.(P66S)	rs140030018	D (0.042)	P (0.948)	24.90	0.0001	0.0000 (0)	0.0052 (2)
RXRA	chr9:137321000	c.666C>T	p.(A222=)	rs137871665	NA	NA	21.00	0.0021	0.0051 (2)	0.0026 (1)
COL5A1	chr9:137582793	c.145C>T	p.(H49Y)	rs372168541	T (0.073)	P (0.837)	24.00	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137591755	c.278C>T	p.(A93V)	rs41306397	D (0.005)	B (0.006)	23.60	0.003	0.0025 (1)	0.0026 (1)
COL5A1	chr9:137591818	c.341C>A	p.(A114D)	rs147589613	D (0.007)	P (0.896)	34.00	0.001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137591844	c.367C>G	p.(Q123E)	rs142114921	T (0.088)	D (0.985)	24.60	0.0003	0.0025 (1)	0.0000 (0)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
COL5A1	chr9:137593038	c.513C>T	p.(S171=)	rs754997436	NA	NA	19.80	0.0000	0.0025 (1)	0.0026 (1)
COL5A1	chr9:137593088	c.563C>G	p.(T188S)	NA	T (0.146)	B (0.370)	24.00	0.0000	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137593122	c.597C>G	p.(I199M)	rs147008954	D (0.002)	P (0.916)	15.30	0.0002	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137593123	c.598G>A	p.(D200N)	rs142890619	D (0.045)	D (1.000)	24.90	0.0006	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137619170	c.713A>G	p.(E238G)	NA	T (0.051)	D (0.998)	27.20	0.0000	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137619196	c.739G>A	p.(A247T)	rs769115550	T (0.396)	P (0.672)	22.40	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137619218	c.761C>T	p.(S254L)	rs144844792	T (0.116)	D (0.981)	25.00	0.0002	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137623967	c.1383C>G	p.(I461M)	rs61736827	D (0.005)	P (0.662)	17.70	0.0000	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137630636	c.1476C>T	p.(V492=)	rs141093527	NA	NA	17.20	<0.0001	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137648614	c.1831C>T	p.(R611W)	rs147329970	D (0.000)	D (1.000)	35.00	<0.0001	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137650103	c.1896C>T	p.(F632=)	rs376478864	NA	NA	15.70	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137657580	c.2088C>T	p.(P696=)	rs146757272	NA	NA	16.00	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137658307	c.2096C>T	p.(T699M)	rs142313124	T (0.071)	P (0.820)	23.90	0.0008	0.0076 (3)	0.0026 (1)
COL5A1	chr9:137666737	c.2364C>A	p.(Y788*)	rs778834633	NA	NA	37.00	0.0003	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137674521	c.2439C>T	p.(D813=)	rs148648778	NA	NA	18.10	0.0006	0.0051 (2)	0.0000 (0)
COL5A1	chr9:137676852	c.2502C>T	p.(P834=)	rs144775947	NA	NA	16.40	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137703220	c.3564C>A	p.(I1188=)	rs766961124	NA	NA	23.20	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137703346	c.3591C>T	p.(D1197=)	rs370349155	NA	NA	17.90	0.0004	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137706723	c.3987C>T	p.(P1329=)	rs770802769	NA	NA	18.70	0.0000	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137708884	c.4135C>T	p.(P1379S)	rs61739195	D (0.006)	D (1.000)	25.80	0.0098	0.0152 (6)	0.0208 (8)
COL5A1	chr9:137710863	c.4410C>T	p.(P1470=)	rs41310953	NA	NA	18.20	0.0021	0.0025 (1)	0.0052 (2)
COL5A1	chr9:137714878	c.4643C>T	p.(S1548L)	rs147398633	D (0.044)	D (0.996)	24.60	<0.0001	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137715269	c.4652C>T	p.(T1551I)	rs863223460	T (0.173)	B (0.002)	23.20	0.0000	0.0000 (0)	0.0026 (1)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
COL5A1	chr9:137715291	c.4674C>T	p.(G1558=)	NA	NA	NA	17.90	0.0000	0.0025 (1)	0.0000 (0)
COL5A1	chr9:137716512	c.4765G>A	p.(A1589T)	rs377138881	T (0.218)	D (0.998)	23.90	<0.0001	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137716653	c.4906G>A	p.(A1636T)	rs113452150	D (0.023)	D (0.995)	27.00	0.0002	0.0000 (0)	0.0026 (1)
COL5A1	chr9:137726950	c.5270C>T	p.(T1757M)	rs2229817	D (0.008)	D (0.995)	25.60	0.0019	0.0000 (0)	0.0052 (2)
COL5A1	chr9:137727015	c.5335A>G	p.(N1779D)	rs780400029	D (0.050)	B (0.183)	23.90	0.0000	0.0000 (0)	0.0026 (1)
FOXO1	chr13:41134313	c.1315A>G	p.(I439V)	rs146471778	D (0.038)	B (0.000)	11.50	<0.0001	0.0000 (0)	0.0026 (1)
FOXO1	chr13:41134320	c.1308G>C	p.(Q436H)	rs767235452	D (0.009)	D (0.990)	24.20	<0.0001	0.0000 (0)	0.0026 (1)
FOXO1	chr13:41134423	c.1205C>T	p.(T402M)	rs148177044	D (0.002)	D (1.000)	25.40	<0.0001	0.0025 (1)	0.0000 (0)
FOXO1	chr13:41134459	c.1169C>T	p.(S390L)	rs756553520	D (0.015)	D (1.000)	26.40	0.0000	0.0000 (0)	0.0026 (1)
RAD51	chr15:41001295	c.416C>T	p.(T139M)	rs148345609	D (0.023)	D (0.993)	34.00	0.0003	0.0025 (1)	0.0000 (0)
RAD51	chr15:41021733	c.675C>T	p.(T225=)	rs147352002	NA	NA	19.00	0.0001	0.0025 (1)	0.0000 (0)
RAD51	chr15:41022106	c.830C>T	p.(A277V)	rs532630164	T (0.080)	P (0.529)	24.60	<0.0001	0.0025 (1)	0.0000 (0)
BANP	chr16:88017786	c.283A>C	p.(K95Q)	rs776932248	D (0.046)	D (1.000)	23.80	<0.0001	0.0025 (1)	0.0000 (0)
BANP	chr16:88017801	c.298A>C	p.(I100L)	rs547536427	T (0.073)	P (0.885)	23.90	0.0003	0.0025 (1)	0.0000 (0)
SLC4A11	chr20:3209320	c.2274G>A	p.(S758=)	rs200879869	NA	NA	17.20	0.0002	0.0000 (0)	0.0026 (1)
SLC4A11	chr20; 3209500	c.2224G>A	p.(G742R)	rs143965185	D (0.001)	D (0.999)	34.00	0.0004	0.0000 (0)	0.0026 (1)
SLC4A11	chr20; 3210069	c.1820T>G	p.(I607S)	rs748984296	D (0.009)	P (0.629)	23.60	0.0002	0.0025 (1)	0.0000 (0)
SLC4A11	chr20; 3210079	c.1810G>A	p.(V604M)	rs771123757	T (0.257)	P (0.924)	12.00	<0.0001	0.0025 (1)	0.0000 (0)
SLC4A11	chr20; 3210278	c.1682C>T	p.(T561M)	rs755379986	T (0.106)	P (0.598)	9.30	0.0000	0.0025 (1)	0.0000 (0)
SLC4A11	chr20; 3211438	c.1270T>C	p.(F424L)	NA	T (0.129)	P (0.951)	22.50	0.0000	0.0025 (1)	0.0000 (0)
SLC4A11	chr20; 3211846	c.1039C>T	p.(R347W)	rs138137682	T (0.188)	B (0.005)	21.60	0.0016	0.0025 (1)	0.0000 (0)
SLC4A11	chr20; 3214281	c.656G>C	p.(C219S)	rs746477170	T (0.053)	P (0.817)	14.10	<0.0001	0.0000 (0)	0.0026 (1)
SLC4A11	chr20; 3214738	c.562C>T	p.(R188W)	rs200372280	T (0.081)	P (0.844)	21.80	<0.0001	0.0000 (0)	0.0026 (1)

Gene	Position	Nucleotide Variant	Protein Variant	Variant ID	SIFT	PolyPhen2	CADD	Alternate allele frequencies		
								ExAC NFE	Controls (AC)	Cases (AC)
SLC4A11	chr20; 3214750	c.550G>A	p.(G184R)	NA	T (0.267)	D (0.998)	23.20	0.0000	0.0025 (1)	0.0000 (0)
VSX1	chr20:25058389	c.740C>G	p.(P247R)	rs576300014	D (0.016)	D (1.000)	33.00	0.001	0.0025 (1)	0.0000 (0)
VSX1	chr20:25058419	c.710T>C	p.(L237P)	rs143704357	D (0.002)	D (1.000)	32.00	0.0000	0.0000 (0)	0.0026 (1)
VSX1	chr20:25058429	c.700G>A	p.(E234K)	NA	D (0.017)	D (1.000)	25.80	0.0000	0.0000 (0)	0.0026 (1)
VSX1	chr20:25060096	c.479G>A	p.(G160D)	rs74315433	T (0.571)	P (0.692)	18.70	0.0033	0.0000 (0)	0.0052 (2)
SOD1	chr21:33040833	c.407C>T	p.(T136I)	rs781031581	T (0.099)	B (0.124)	17.60	<0.0001	0.0025 (1)	0.0000 (0)

Variant ID = as available from the dbSNP147 database.
CADD = the scaled CADD score.
ExAC NFE = the alternate allele frequency observed in the non-Finnish European population of the Exome Aggregation Consortium (ExAC) database.
AC = the alternate allele count in the given cohort.
SIFT and Polyphen2 classifications: D = deleterious/damaging; P = possibly damaging; T = tolerated; B = benign.

Table 3.5 – Burden test results for genes in which at least one potentially pathogenic variant was identified in the case cohort.

	Case Alleles		Control Alleles			
Gene	**Alt.**	**WT**	**Alt.**	**WT**	**P**	**OR [95% CI]**
CAST	6	764	7	785	> 0.99*	0.88 [0.26-2.92]
COL4A3	21	749	39	753	0.02*	0.54 [0.31-0.96]
COL4A4	7	763	9	783	0.85*	0.80 [0.27-2.35]
COL5A1	32	738	22	770	0.18*	1.52 [0.85-2.73]
FNDC3B	5	765	2	790	0.28	2.58 [0.45-19.24]
FOXO1	3	767	1	791	0.37	3.09 [0.29-77.33]
HGF	3	767	6	786	0.51	0.51 [0.10-2.30]
IL1A	3	767	8	784	0.23	0.38 [0.08-1.59]
IMMP2L	1	769	3	789	0.63	0.34 [0.01-3.67]
MPDZ	40	730	62	730	0.05*	0.65 [0.42-0.99]
NFIB	3	767	3	789	> 0.99	1.03 [0.17-6.38]
RAB3GAP1	5	765	5	787	> 0.99	1.03 [0.26-4.10]
RXRA	1	769	2	790	> 0.99	0.51 [0.02-7.20]
SLC4A11	4	766	6	786	0.75	0.68 [0.16-2.74]
TF	8	762	9	783	> 0.99*	0.91 [0.32-2.59]
VSX1	4	766	1	791	0.21	4.13 [0.44-97.26]

Alt. = the number of alternate alleles (ie. the number of alleles that carry a potentially pathogenic variant).
WT = the number of wild type (reference) alleles (ie. the number of alleles that do not carry a potentially pathogenic variant).
P = p-value.
OR [95% CI] = odds ratio and the corresponding 95% confidence interval.
* P-values obtained using a Yates corrected chi-square test. All other p-values were obtained using a Fisher's exact test. P-values <0.0024 were considered significant.

3.4.4 Potentially pathogenic variants in *ZNF469*

A total of 385 cases, 346 population controls and 230 screened controls were included in the following analysis. Demographic details are provided in Table 3.6.

Table 3.6 – Demographics of the keratoconus cases and control groups used in the analysis of *ZNF469*.

Cohort	n	Mean age* (range)	% Female	Disease status
Cases	385	45.2 (14 – 85)	44.2	Affected
Population controls^	346	69.5 (46 – 86)	100	Unscreened
Screened Controls	230	68.6 (8 – 92)	56.5	Unaffected

n = the number of individuals.
* Age is reported in years.
^ Age was collected at bone densitometry. For other groups, age was recorded at blood draw.

A stringent coverage filter was applied to all sequence datasets and therefore regions containing the majority of the first exon of *ZNF469*, the beginning of the second exon, and the regions encoding the zinc finger domains were excluded from analyses that compared the cases to the population control cohort (Figure 3.1). Comparisons using the population controls included a total of 6700 bases across *ZNF469* with high quality sequence data, corresponding to 56.5% of the coding regions. In contrast, a total of 11448 bases (96.5% of the coding regions) obtained sufficient coverage across the gene in both the cases and screened controls. For comparisons using the screened controls, only two small regions (85 bp and 279 bp) either side of the intron were excluded from analysis due to poor coverage (Figure 3.1). By removing these regions from analysis, the single region that did not meet the minimum depth threshold in the pooled targeted gene screen data was also excluded. In 34 of the 44 DNA pools a 70 bp region (chr16:88496809-88496879) corresponding to the distal end of exon 1 did not reach the minimum depth threshold for inclusion. Of the remaining DNA pools, one had a 140 bp region containing bases below threshold (chr16:88496809-88496949), three DNA pools had insufficient coverage for a 2 bp region (chr16:88496809-88496810), and six DNA pools had sufficient coverage across the entirety of exon 1. A summary of the coverage metrics for the pooled targeted gene screen dataset is presented in Table 3.7.

Table 3.7 – A summary of the coverage metrics for *ZNF469* for the pooled targeted gene screen dataset as reported by Agilent Technologies, averaged across all DNA pools.

Covered region	Position (hg19)	Mean depth (range)	Mean min. depth (range)	Mean max. depth (range)	Mean bases with depth <10 (range)
1	chr16:88493639-88494052	1,676 (713-2,255)	282 (113-452)	5,058 (2,249-7,176)	0
2	chr16:88494053-88502070	2,559 (1,149-3,623)	10 (0-32)	8090 (3,791-12,249)	57 (0-140)
3	chr16:88502074-88502132	1082 (485-1,897)	1,054 (472-1,865)	1,890 (871-2,822)	0
4	chr16:88502141-88503852	2,627 (1,168-3,749)	120 (45-197)	5,705 (2,519-8,305)	0
5	chr16:88503862-88505962	3,219 (1,475-4,570)	132 (46-227)	12,993 (6,381-19,874)	0

Min. = minimum
Max. = maximum

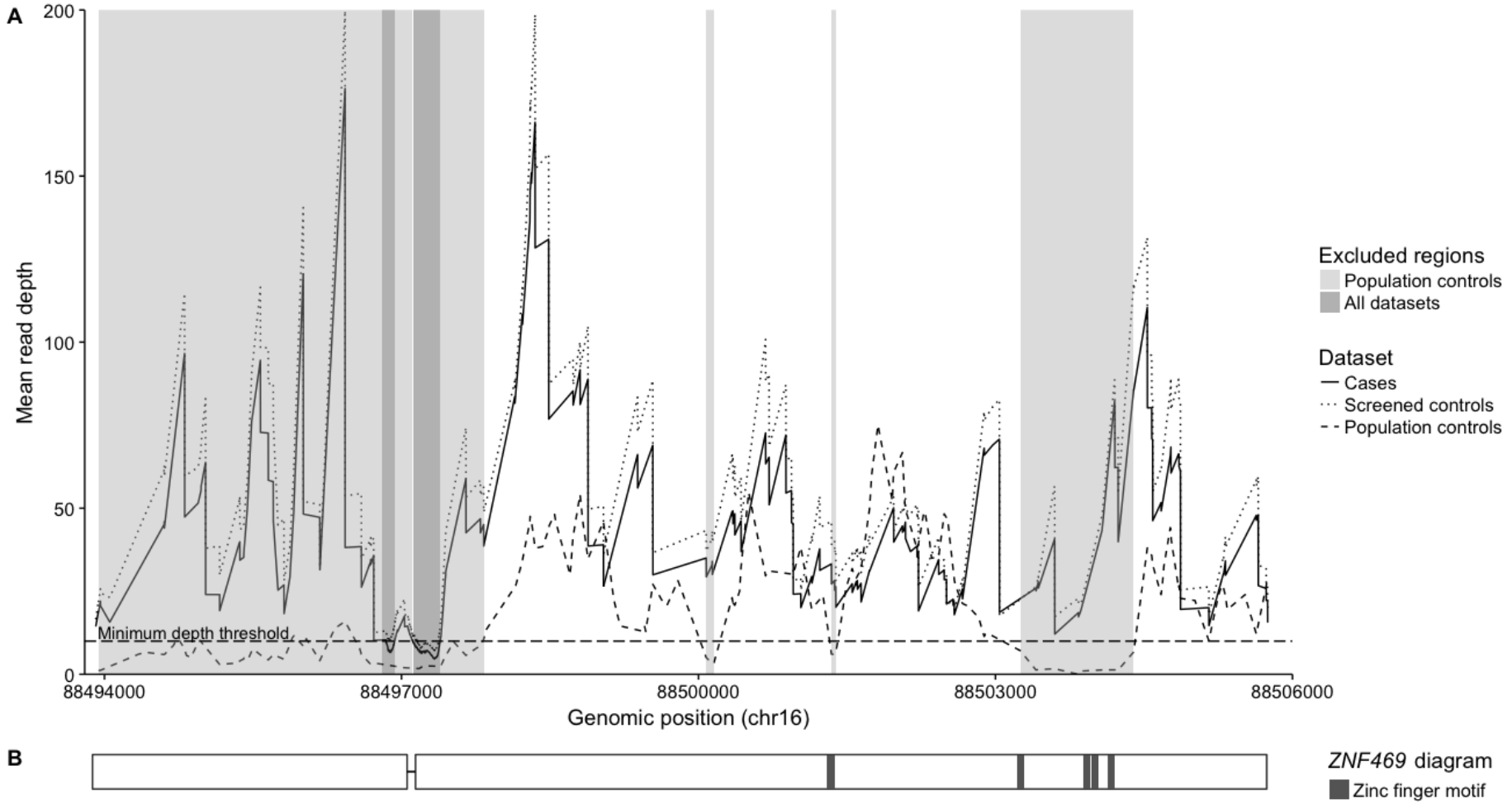


Figure 3.1 – Sequencing coverage in the WES datasets across *ZNF469*.
(A) Depicts the coverage across *ZNF469* in each dataset, based on the mean read depth at variant positions. The horizontal dashed line indicates the minimum depth threshold accepted (10 reads). Regions where the mean read depth in the population controls was below this threshold are shaded light grey and were excluded from analyses in which these controls were used. Regions with insufficient coverage in all cohorts are shaded dark grey. (B) A schematic of *ZNF469* where dark grey boxes indicate the position of the zinc finger motifs and the small intron is indicated by the horizontal line.

Across all cohorts, 49 variants (46 single nucleotide variants and three small deletions) fulfilled the criteria for rare potentially pathogenic variants in *ZNF469* under Filtering Strategy 1 (Table 3.8). Of these, nineteen variants were observed in cases only, while 17 were unique to the control cohorts and the remaining 13 were in both cases and controls. Eleven variants had been observed in previous studies in either cases or controls, with an additional two variants that are not identical, but occur at the same amino acid as a previously reported variant. A total of 33 rare potentially pathogenic variants were observed 73 times in cases and 23 variants were called 48 times in the screened controls, which was not significantly different ($p = 0.66$, Table 3.9). When considering the comparison between the cases and the population controls, 17 variants were identified 31 times in cases and 10 variants were observed 22 times in the controls, with no significant difference between groups ($p = 0.47$, Table 3.9).

Filtering Strategy 2 identified 33 very rare potentially pathogenic variants in *ZNF469* across all cohorts (Table 3.7). Sixteen of these variants were unique to keratoconus patients and two, p.(C1693F) and p.(P2548L), were identified in two cases each. Fifteen variants were observed in the control cohorts only, including p.(P3372L), which was identified in two individuals. Two variants, p.(P626_G628del) and p.(E3781K), were identified in both a case and a control. All other variants were observed in a single individual. Two variants, p.(S2242Y) and p.(P3372L), were previously observed in the study by Lechner *et al.*[145] and p.(E935G) is located at the same amino acid as a variant identified in the study by Davidson and colleagues.[147] In total, 18 variants were identified 20 times in cases and 12 variants were observed 13 times in the screened controls. For the population control comparison, 10 variants were observed in 12 cases and 5 variants were identified in the controls. Similar to the results for rare variants, very rare variants were not enriched in cases compared to either the screened controls ($p = 0.96$) or the population controls ($p = 0.15$, Table 3.8). Furthermore, the SKAT analyses demonstrated no significant enrichment of variants predicted to be damaging ($p = 0.06$, Table 3.10).

For the power calculations, the high-risk allele frequency was assumed to be the frequency of variants identified by the filtering strategies in the 230 screened controls. This was determined to be 0.104 (48/460) for Strategy 1 and 0.028 (13/460) for Strategy 2. Using these values, the present analysis of *ZNF469* had 80% power to detect a relative risk of 1.5 and 2.0, respectively.

Table 3.8 – Rare potentially pathogenic variants in *ZNF469* variants included in analysis under Filtering Strategy 1

Position (hg19)	Nucleotide Variant	Protein Variant	PhastCons	Alternate allele frequencies				FS2	Ref
				ExAC NFE	Cases (AC)	Screened Controls (AC)	Population Controls (AC)		
chr16:88494603	c.725G>T	p.(S242I)	0.9930	0.0024	0.0013 (1)	0.0022 (1)	*	N	
chr16:88494809	c.931G>A	p.(G311R)	0.0000	0.0000	0.0013 (1)	0.0000 (0)	*	Y	
chr16:88495361	c.1483C>T	p.(P495S)	0.0040	0.0015	0.0013 (1)	0.0022 (1)	*	N	
chr16:88495385	c.1507C>T	p.(R503W)	0.9980	0.0003	0.0013 (1)	0.0000 (0)	*	Y	
chr16:88495400	c.1522G>A	p.(A508T)	0.0000	0.0003	0.0013 (1)	0.0000 (0)	*	Y	
chr16:88495443	c.1565G>T	p.(G522V)	0.0010	0.0000	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88495461	c.1583C>G	p.(P528R)	0.0000	0.0000	0.0013 (1)	0.0000 (0)	*	Y	
chr16:88495487	c.1609G>A	p.(V537M)	0.0000	0.0010	0.0038 (3)	0.0022 (1)	*	N	145
chr16:88495568	c.1690G>C	p.(G564R)	0.3100	0.0002	0.0013 (1)	0.0000 (0)	*	Y	
chr16:88495575	c.1697C>T	p.(A566V)	0.0000	0.0098	0.0115 (9)	0.0283 (13)	*	N	146, 147
chr16:88495753	c.1875_1883del	p.(P626_P628del)	0.0000	0.0000	0.0013 (1)	0.0022 (1)	*	Y	
chr16:88495872	c.1994C>T	p.(P665L)	0.0420	0.0062	0.0102 (8)	0.0109 (5)	*	N	145
chr16:88495913	c.2035G>A	p.(E679K)	0.6550	0.0064	0.0013 (1)	0.0022 (1)	*	N	147
chr16:88496430	c.2552T>C	p.(M851T)	1.0000	0.0000	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88496682	c.2804A>G	p.(E935G)	0.0040	0.0000	0.0013 (1)	0.0000 (0)	*	Y	147†
chr16:88497443	c.3481C>A	p.(P1161T)	0.0000	0.0000	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88497830	c.3868G>A	p.(D1290N)	0.0020	0.0000	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88498350	c.4388C>T	p.(T1463M)	0.0020	0.0035	0.0000 (0)	0.0022 (1)	0.0058 (4)	N	145
chr16:88499040	c.5078G>T	p.(C1693F)	0.0000	0.0002	0.0026 (2)	0.0000 (0)	0.0000 (0)	Y	
chr16:88500687	c.6725C>A	p.(S2242Y)	0.0000	0.0001	0.0000 (0)	0.0000 (0)	0.0014 (1)	Y	145
chr16:88500713	c.6751C>T	p.(P2251S)	0.0040	0.0000	0.0000 (0)	0.0022 (1)	0.0000 (0)	Y	

Position (hg19)	Nucleotide Variant	Protein Variant	PhastCons	Alternate allele frequencies				FS2	Ref
				ExAC NFE	Cases (AC)	Screened Controls (AC)	Population Controls (AC)		
chr16:88500822	c.6860C>G	p.(P2287R)	0.0010	0.0002	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	
chr16:88500957	c.6995C>T	p.(P2332L)	0.0000	0.0033	0.0064 (5)	0.0000 (0)	0.0000 (0)	N	
chr16:88501145	c.7183C>A	p.(P2395T)	0.0000	0.0044	0.0038 (3)	0.0022 (1)	0.0000 (0)	N	
chr16:88501224	c.7262G>A	p.(R2421H)	0.0000	0.0002	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	
chr16:88501344	c.7382G>A	p.(R2461Q)	1.0000	0.0002	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88501431	c.7469C>A	p.(P2490H)	0.0000	0.0034	0.0013 (1)	0.0000 (0)	0.0000 (0)	N	145
chr16:88501605	c.7643C>T	p.(P2548L)	0.0000	0.0000	0.0026 (2)	0.0000 (0)	0.0000 (0)	Y	
chr16:88501749	c.7787C>T	p.(P2596L)	0.0000	0.0010	0.0000 (0)	0.0000 (0)	0.0058 (4)	N	
chr16:88501995	c.8033C>T	p.(A2678V)	0.0000	0.0000	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	
chr16:88502222	c.8260C>T	p.(H2754Y)	0.0000	0.0009	0.0000 (0)	0.0022 (1)	0.0000 (0)	Y	
chr16:88502276	c.8314C>A	p.(L2772M)	0.0000	0.0000	0.0000 (0)	0.0000 (0)	0.0014 (1)	Y	
chr16:88502304	c.8342C>T	p.(P2781L)	0.0000	0.0000	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	
chr16:88502583	c.8621C>T	p.(T2874M)	0.0000	0.0011	0.0026 (2)	0.0000 (0)	0.0000 (0)	N	
chr16:88502666	c.8704G>T	p.(D2902Y)	0.9700	0.0010	0.0038 (3)	0.0000 (0)	0.0029 (2)	N	145
chr16:88502723	c.8761C>G	p.(P2921A)	0.0030	0.0000	0.0000 (0)	0.0000 (0)	0.0014 (1)	Y	
chr16:88502862	c.8900C>T	p.(A2967V)	0.0030	0.0000	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	
chr16:88502972	c.9010_9024del	p.(L3004_T3008del)	0.0000	0.0043	0.0051 (4)	0.0022 (1)	0.0014 (1)	N	147†,
chr16:88503039	c.9077A>C	p.(E3026A)	0.9690	0.0000	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	
chr16:88503372	c.9410A>G	p.(E3137G)	1.0000	0.0000	0.0013 (1)	0.0000 (0)	*	Y	
chr16:88503596	c.9634A>T	p.(R3212W)	0.0000	0.0000	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88503838	c.9876G>T	p.(E3292D)	0.0010	0.0000	0.0000 (0)	0.0022 (1)	*	Y	
chr16:88504077	c.10115C>T	p.(P3372L)	1.0000	0.0000	0.0000 (0)	0.0043 (2)	*	Y	145
chr16:88504239	c.10277G>A	p.(R3426Q)	0.3970	0.0078	0.0140 (11)	0.0196 (9)	*	N	145, 147

Position (hg19)	Nucleotide Variant	Protein Variant	PhastCons	Alternate allele frequencies				FS2	Ref
				ExAC NFE	Cases (AC)	Screened Controls (AC)	Population Controls (AC)		
chr16:88504492	c.10530delC	p.(I3510fs)	0.0000	0.0000	0.0000 (0)	0.0000 (0)	0.0014 (1)	Y	
chr16:88504766	c.10804C>T	p.(R3602C)	0.0000	0.0031	0.0013 (1)	0.0022 (1)	0.0087 (6)	N	[145]
chr16:88504775	c.10813T>G	p.(C3605G)	0.0000	0.0000	0.0000 (0)	0.0000 (0)	0.0014 (1)	Y	
chr16:88505303	c.11341G>A	p.(E3781K)	1.0000	0.0005	0.0013 (1)	0.0022 (1)	0.0000 (0)	Y	
chr16:88505654	c.11692G>A	p.(E3898K)	1.0000	0.0000	0.0013 (1)	0.0000 (0)	0.0000 (0)	Y	

Position = the position of SNPs or the start position of insertions or deletions (based on hg19).
PhastCons = the 100-way vertebrate PhastCons score (which ranges from 0 to 1, with 1 being highly conserved).
ExAC NFE = the alternate allele frequency observed in the non-Finnish European population of the Exome Aggregation Consortium (ExAC) database.
AC = the alternate allele count in the given cohort.
FS2 = indicates whether or not the variant was included in Filtering Strategy 2, where Y = yes and N = no.
Ref = References of previous studies if the variant had been previously reported in keratoconus.
* = variant with insufficient coverage.
† = variant previously reported at the same amino acid, but a different nucleotide.

Table 3.9 – Association analyses using chi square or Fisher's exact test under each filtering strategy used for *ZNF469*.

Filtering Strategy 1	Cohort	Alt.	WT	AAF	P	OR [95%CI]
Comparison 1	cases	73	697	0.09	0.66	0.90 [0.60-1.34]
	screened controls	48	412	0.10		
Comparison 2	cases	31	739	0.04	0.47	1.28 [0.71-2.31]
	population controls	22	670	0.03		
Filtering Strategy 2	**Cohort**	**Alt.**	**WT**	**AAF**	**P**	**OR [95%CI]**
Comparison 1	cases	20	750	0.03	0.96	0.92 [0.43-1.97]
	screened controls	13	447	0.03		
Comparison 2	cases	12	758	0.02	0.15*	2.18 [0.71-7.11]
	population controls	5	687	0.01		

Alt. = the number of alternate alleles (ie. the number of alleles that carry a potentially pathogenic variant).
WT = the number of wild type (reference) alleles (ie. the number of alleles that do not carry a potentially pathogenic variant).
AAF = the alternate alleles frequency (Alt./total alleles).
P = p-value obtained using a Yates corrected chi-square test.
*Fisher's exact p-value.

Table 3.10 – SKAT results for *ZNF469*.

Comparison	Alt.	MAC	Carriers	P
Cases vs screened controls	38	415	271	0.06
Cases vs population controls	20	473	337	0.06

Alt. = the number of alternate alleles (ie. the number of variants observed).
MAC = the total number of minor alleles observed within the cohorts (ie. the number of times the minor allele was observed for all variants included in the analysis).
Carriers = the number of individuals carrying the alternate alleles.
P = p-value.

For each cohort, the allele frequencies were plotted against the variant position and mapped to a schematic of the gene (Figure 3.2). The first exon showed a similar pattern of variation between the cases and screened controls. Due to insufficient coverage of this region in the population controls, variation in this group could not be assessed. Of the 13 variants identified in cases in exon 1, two variants, p.(S242I) and p.(R503W), were located at highly conserved nucleotides (PhastCons score > 0.99) and p.(E679K) was at a relatively conserved position with a PhastCons score of 0.655. The p.(R503W) variant was observed in a single case (frequency of 0.0013), whereas both p.(S242I) and p.(E679K) were identified in one case and one screened control (frequency of 0.0022).

Few variants were located in the proximal region of exon 2, with only one variant identified in cases. In contrast, a cluster of variants was observed in cases in the distal half of exon 2 (Figure 3.2). Specifically, variants were observed in the cases in the region spanning just proximally to the first zinc finger motif to midway between the second and third zinc finger motif, with noticeably fewer variants were observed in the control cohorts within the same region. Three variants identified in cases within this cluster were highly conserved with PhastCons scores greater than 0.96: p.(D2902Y), p.(E3026A) and p.(E3137G). The p.(D2902Y) variant was present at a similar frequency in cases (0.0038) and the population controls (0.0029), while p.(E3026A) and p.(E3137G) were present in one case and absent in controls. Two additional variants, p.(E3781K) and p.(E3898K), were identified at highly conserved residues at the very distal end of exon 2 (PhastCons score = 1). These variants were both identified a single case and neither of these variants were observed in the population controls despite good coverage in this region, however, the p.(E3781K) variant was observed in a one screened control.

Across *ZNF469*, three variants – p.(A566V), p.(P665L), and p.(R3426Q) – were identified at a frequency > 0.01 in both the cases and the screened control cohort, despite a MAF < 0.01 in the public ExAC NFE database. The frequency of these variants could not be assessed in the population controls due to insufficient coverage, however all three variants were observed in cases at a similar or lower frequency than the screened controls.

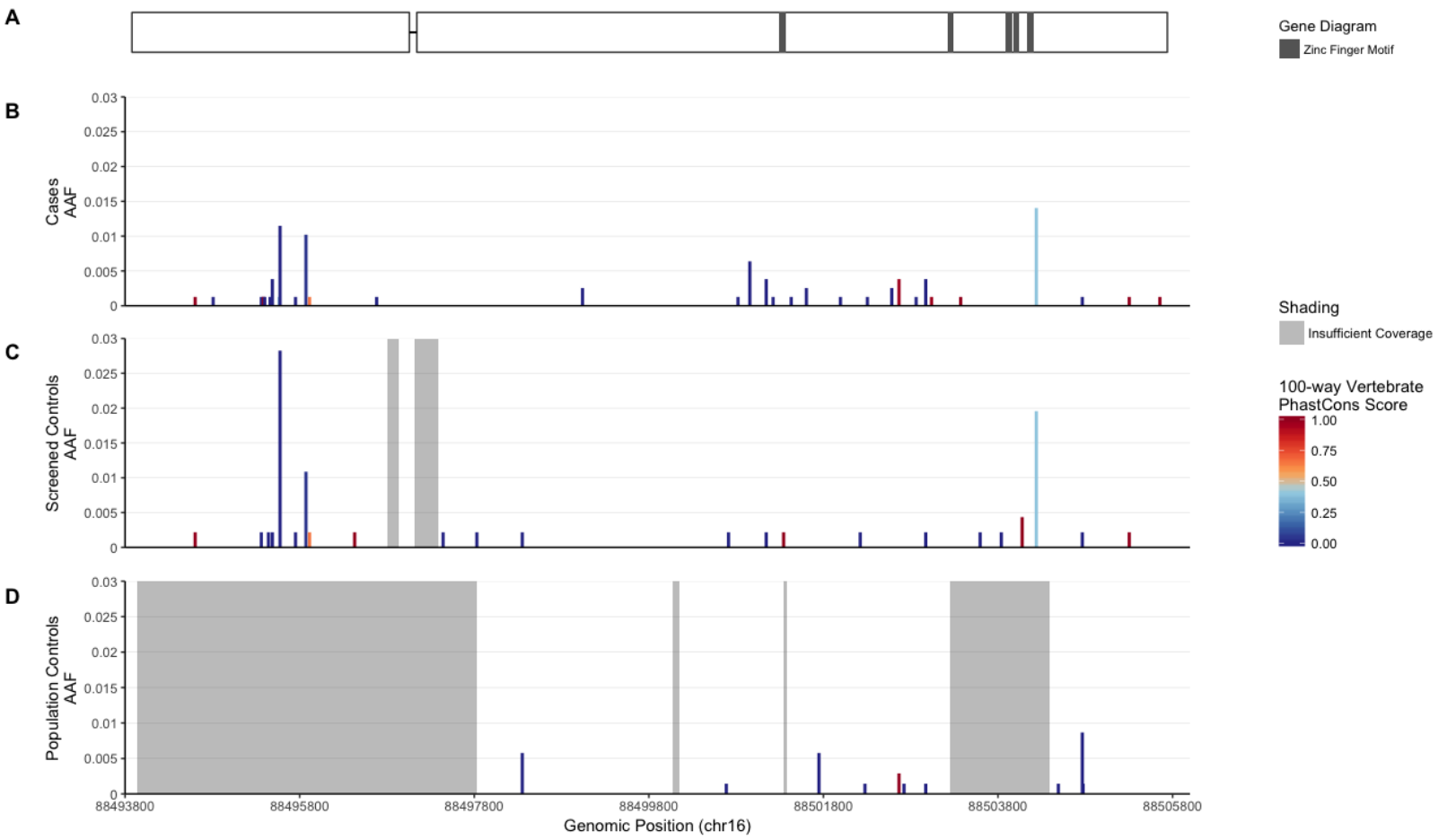


Figure 3.2 – Summary of variants indentified in *ZNF469*.

(A) Depicts a schematic of *ZNF469* with zinc finger motifs shaded dark grey and (B-D) presents bar plots indicating the position and alternate allele frequency (AAF) of rare potentially pathogenic variants in 784 case alleles, 460 screened control alleles and 692 population control alleles respectively. The bars are coloured according to the 100-way vertebrate PhastCons Score for the corresponding position, where dark blue is a score of 0 and red is a score of 1. Light grey shading on the plots for the control cohorts indicate regions with insufficient coverage for variant calling that were excluded from analysis when using these data. The gene schematic and all graphs are aligned vertically to share the same x-axis to allow for comparison.

3.5 DISCUSSION

This study demonstrates that rare potentially pathogenic variants in 22 candidate genes were not enriched in keratoconus in our large cohort of Australians of European descent. A total of 213 rare potentially pathogenic variants were identified across cases and controls in 21 of the 22 genes, however, variants fulfilling these criteria were equally common between cases and controls. No potentially pathogenic variants were identified in *IL1RN* in either cases or controls, suggesting that this gene is highly conserved. Additionally, no potentially pathogenic variants were observed in our cases in *BANP*, *IL1B*, *RAD51* or *SOD1*. Based on these findings, we suggest that rare protein-coding variants that are predicted to be potentially pathogenic within the 22 genes assessed do not contribute broadly to keratoconus development in our cohort.

This study is the largest and most comprehensive study of keratoconus candidate genes to date, including 22 genes and combining WES data and targeted gene sequencing using pooled DNA samples. Following extensive validation experiments, we demonstrated a high level of consistency of variant calls for individuals sequenced by both methods, validating the utility of pooling DNA samples to maximise cost-effectiveness. Furthermore, these validation experiments were used to develop stringent thresholds that were then applied to remaining data to ensure only high-quality variants were included in analyses. While variants were not assessed for validation in the control cohorts, the inclusion of additional control variants that were filtered out due to the stringent quality and depth thresholds would only strengthen our findings of no association. Furthermore, gaps in coverage, occurring either at the probe design stage or at the sequencing phase, were a limitation of this study. Similarly, the capture methods were specifically designed to capture protein-coding regions, and although non-coding variants were not excluded, very few were observed. In addition, for all genes except *ZNF469*, insertion and deletion variants were not assessed due to the challenges of calling such variants from the targeted sequencing data generated using pooled DNA. These limitations mean that some variants will have been missed, but this is not expected to be a major bias between cases and controls due to limiting the analysis to regions adequately covered by all methods. Some specific variants that have previously been reported in keratoconus could not be assessed. This is particularly important for *SOD1*, as the variant previously associated with keratoconus was a 7 bp intronic deletion, outside of the capture regions in this study. Similarly, *IL1RN* and *SLC4A11* were implicated by an intronic SNP and a 54 bp intronic deletion (respectively) which almost completely co-segregated with keratoconus in an Ecuadorian family.[66] Although these specific variants could not be assessed in our study, the overarching design and aim was to examine the protein-coding regions of the selected candidate genes for enrichment of potentially pathogenic variants in keratoconus.

For the few genes that have previously been investigated in keratoconus cohorts, *in silico* tools that predict the pathogenicity of variants such as PolyPhen2 and SIFT have been used to help differentiate

between benign variants and potentially pathogenic variants. These algorithms, however, can only assess nonsynonymous variants. These two tools were used to classify variants as potentially pathogenic in *ZNF469*, allowing for direct comparison of the results from the present study and those previously published, however, for all other genes, scaled CADD scores were also used to aid the classification of these variants. CADD uses a machine learning method to predict the deleteriousness of variants such that a scaled score above 10 refers to the top 10% of variants ranked by deleteriousness, a scaled score of 20 or above includes the top 1% of variants and so on.[134] Unlike SIFT and PolyPhen2, CADD also scores synonymous and non-coding variants, allowing for the inclusion of these types of variants in the present study. A scaled CADD score of 15 was selected as the minimum threshold for synonymous and non-coding variants. Nonsynonymous variants with a MAF < 0.01 were included if they were predicted to be damaging by SIFT; or damaging/possibly or probably damaging by PolyPhen2; or obtained a CADD scaled score of ≥ 15. This broad definition of potentially pathogenic variants was designed to minimise the exclusion of likely important disease-related variants.

This study compared the frequency of potentially pathogenic variants between Australian keratoconus cases and controls, using a control cohort consisting of 396 females. These individuals were not screened for keratoconus, however, considering the prevalence of keratoconus in Caucasians is between approximately 1 in 375 and 1 in 2000[4, 6, 178] it is unlikely that more than one or two individuals in this cohort have keratoconus, if any. Considering the large sample size for both the case and control cohorts, this is unlikely to affect our findings. Additionally, the controls were 100% female while 44.2% of the cases were female. This is a potential limitation, however, while epidemiological studies report a slightly higher prevalence in males compared to females in Caucasian populations, these differences are not significant.[4, 6, 178] Moreover, all of the genes assessed in the present study are autosomal, making sex-based differences in the frequency of variants unlikely. Consequently, the use of this all-female cohort is unlikely to affect the outcomes of the present study. Due to poor coverage of *ZNF469*, particularly in the WES data for the population controls, a screened control cohort consisting of 230 individuals was also used for analysis. These individuals were almost all affected by advanced glaucoma. However, the GWAS that identified the association between the SNP upstream of *ZNF469* (rs9938149) and CCT as well as keratoconus, also assessed this SNP for association with glaucoma and demonstrated no association.[87] Based on this, we do not expect the use of this cohort to confound our findings.

VSX1, which encodes a vertebrate paired-like homeodomain transcription factor with known ocular expression,[179, 180] is the most studied gene in keratoconus. It was initially studied as a candidate gene for posterior polymorphous corneal dystrophy (PPCD; OMIM 122000) as it was located within a linkage region for this disease.[89, 181] PPCD is a rare, bilateral corneal dystrophy that primarily affects the endothelium and results in variable degrees of visual impairment and has been associated with keratoconus.[182-186] It was therefore hypothesised that the two diseases may share a common genetic

basis. The original paper identified four *VSX1* variants in keratoconus cases that were absent in 277 controls,[89] as well as p.(G160D) and p.(P247R) in a family with PPCD.[89] Subsequently, the p.(G160D) variant has been identified in keratoconus cases in two Italian studies[90, 96] a European cohort,[94] and in two cases in the current study. This was the only variant observed in our cases that had been reported in other keratoconus cases. In contrast, the p.(G160D) variant was found at similar frequencies in both cases and controls in a Han Chinese cohort,[97] indicating that this variant is not highly penetrant for keratoconus, at least in the Chinese population. Interestingly, the p.(G160V) variant, which results in a different amino acid substitution at the same position, has been observed in cases in two Korean studies.[92, 156] The p.(P247R) variant originally reported in the PPCD family[89] and subsequently reported in keratoconus[90, 96, 100] was observed in a single control subject in our study. As our controls were not screened for eye disease it is possible that a small number of individuals in this group may have keratoconus or PPCD and therefore the involvement of this variant in disease cannot be ruled out. Furthermore, as keratoconus is a complex disease it is likely that unaffected individuals may carry risk alleles, without ever developing disease, therefore, to assess the potential role of specific variants such as p.(P247R), a large meta-analysis is required. Additionally, p.(L237P) and p.(E234K) were identified for the first time in a keratoconus cohort, each observed in one case. Taken together, the *VSX1* gene may contribute in a very small number of cases with clear segregation in families identified,[90, 91, 95, 98] however, rare potentially pathogenic variants in this gene do not contribute widely to keratoconus susceptibility in our cohort.

The present analysis of *ZNF469* included more keratoconus cases than the combined number of cases previously studied.[145-148] This allowed the location of the rare potentially pathogenic variants to be mapped across the gene in cases and both control cohorts, demonstrating that rare potentially pathogenic variants span the whole gene, with particular aggregation in the first exon and the distal half of the second exon. Only eight variants observed in cases were located at highly conserved nucleotides. Four of these variants were identified at similar frequencies in both cases and controls and four were only observed in a single case. It is possible that some of the rare protein-coding variants may contribute to keratoconus susceptibility in these few cases, and based on our power calculations it is feasible that our study was underpowered to detect an association of variants with small odds ratios ($OR \leq 1.5$ for rare variants and $OR \leq 2$ for very rare variants), but on the whole, the evidence indicates that rare, potentially pathogenic variants in *ZNF469* do not make a substantial contribution to keratoconus risk. This finding is consistent with the work by Davidson and colleagues,[147] which showed that uncommon variants ($MAF < 0.025$) did not segregate with disease in families with keratoconus and therefore, at least in isolation, do not contribute to keratoconus susceptibility. Furthermore, the results of a Polish study[148] indicate that potentially pathogenic variants are not enriched in keratoconus, and that *ZNF469* is highly allelic in the general population.

The first reports to investigate coding variants in *ZNF469* were conducted in small cohorts from UK/Swiss and New Zealand populations, published by Lechner *et al.*[145] and Vincent *et al.*[146] respectively. Subsequent studies, including this one, have used similar filtering strategies to allow for direct comparison. Our Filtering Strategy 1 was based on the criteria used in Vincent *et al.*, while Filtering Strategy 2 was based on the method used by Lechner *et al.* with two key changes. Firstly, Lechner *et al.* removed any variants from analysis that were present in both their cases and controls. We did not do this as keratoconus is a complex disease and therefore it is likely (and expected) that unaffected individuals will carry risk associated variants without ever developing disease. Secondly, Lechner *et al.* only used SIFT predictions to classify variants as potentially pathogenic. SIFT uses protein conservation to calculate pathogenicity by comparing a query sequence to similar sequences with similar function.[132] As *ZNF469* is a highly variable gene with low conservation in lower mammals and vertbrates,[187] SIFT may misclassify deleterious variants in regions of low conservation. In contrast, PolyPhen2 uses the properties of the substituted amino acids and the proximity to functional domains or structural features, as well as protein conservation to predict pathogenicity.[133] As the structure and function of the ZNF469 protein remains largely unknown,[187] PolyPhen2 is likely to better assess regions with poor conservation, particularly the regions that flank the zinc finger domains. Therefore, our study used both SIFT and PolyPhen2 to better capture the pathogenicity of nonsynonymous variants identified in *ZNF469*. Our study used these robust and complementary methods to replicate the analytical strategies of the previous studies; however, our findings do not support any enrichment of rare potentially pathogenic variants in *ZNF469* in keratoconus cases.

As suggested by Davidson *et al.*[147] it is likely that variation in *ZNF469* is under-represented in public databases as a result of the poor coverage of the gene by the older WES capture techniques. According to the ExAC Browser the mean coverage for *ZNF469* (ENSG00000225614) is 7.3 reads and the proportion of individuals with at least 10X coverage is less than 20%. The poor coverage of *ZNF469* in the population control dataset, as well as the occurrence of three variants that were identified in our cases and screened controls at a frequency of > 0.01, despite being annotated with a MAF < 0.01 in the ExAC NFE database, supports this. As one might expect, these variants could not be assessed in the population controls due to insufficient coverage. All three of these variants were located at relatively non-conserved nucleotides (PhastCons scores < 0.4). Two of these variants, p.(P665L) and p.(R3426Q) were observed at similar frequencies in both our cases and controls and therefore were hypothesised to be benign polymorphisms. These variants were similarly classified in the study by Lechner and colleagues, but were not identified in other reports.[145-148] The third variant, p.(A566V), was identified at more than twice the frequency in the screened controls (0.028) than the cases (0.012). Vincent *et al.*[146] reported this variant in one Indian and two Caucasian keratoconus cases, while Davidson *et al.*[147] identified the variant in two cases and two unaffected individuals from two separate consanguineous families of Middle Eastern origin. In addition, the work by Lechner *et al.*[145] excluded the variant from

analysis due to a MAF > 0.01 in their control cohort. Therefore, we propose that these variants are benign, uncommon polymorphisms. Overall, this work demonstrates that the previously reported large effects of rare variants in *ZNF469* are likely spurious and are brought about by biased reporting of variants.

3.6 CONCLUSION

This study demonstrated that the overall the frequency of potentially pathogenic variants was not different between cases and controls in 22 candidate genes in our large cohort of Australians of European descent. The included genes were all literature-based candidates, proposed to play a role in disease based on their proximity to GWAS hits or because they map to linkage regions identified in family studies, as well as, the function of the encoded protein. While specific rare protein-coding variants in these genes may contribute to keratoconus-risk in a small proportion of cases, the work presented in this chapter suggests that they do not contribute to disease in the vast majority of keratoconus patients. Perhaps some of these genes do contribute to keratoconus susceptibility, but via an alternative mechanism of disease such as altered regulation and expression. Fine-mapping and re-sequencing techniques are required to identify the functional risk-associated variants at keratoconus-associated loci, determine the mechanism of disease and aid our understanding of key biological pathways involved in keratoconus. Furthermore, as demonstrated by the success of identifying the keratoconus-associated gene *mir184*,[65, 77, 143] family studies paired with massively-parallel sequencing technologies may be a powerful method for elucidating specific disease-causing variants. Together, these methods would allow for less biased approaches for variant identification and candidate gene selection without a prior hypothesis. Overall, these findings do not support the overarching hypothesis of this study, that rare protein-coding variants contribute to keratoconus development, and instead suggest that alternative hypotheses should be explored.

CHAPTER 4: IDENTIFIYING PUTATIVELY DISEASE-CASUSING VARIANTS IN FAMILIES WITH MULTIPLE CASES OF KERATOCONUS

4.1 INTRODUCTION

Studying families enriched for disease is a powerful strategy for identifying rare, disease-causing variants. Linkage analysis is a statistical method that has been used for decades to identify regions of the genome that segregate with disease through a family. Parametric linkage analysis is used when the inheritance pattern of the disease or trait of interest is known as it applies a specific disease model. Conversely, non-parametric linkage analysis does not make any assumptions about the disease model and instead tests for increased allele sharing between affected individuals. This method is applied when the inheritance pattern is less clear. Furthermore, homozygosity mapping is the most robust discovery strategy for identifying the disease-causing variant in families with recessive disease and a recent history of consanguinity. This method identifies alleles that are homozygous identical-by-descent in affected individuals. Traditionally, these disease mapping methods were conducted using a relatively small number of polymorphic microsatellite markers across the genome and the identified regions of interest were then fine-mapped using a denser selection of markers surrounding the region. More recently, single nucleotide polymorphism (SNP) arrays that genotype hundreds of thousands of SNPs across the genome have become the norm, reducing the need for fine-mapping. The regions of interest are then sequenced, traditionally using direct sequencing methods to sequence the coding regions of highly prioritised genes, to identify the causative variants.

In families, keratoconus is often inherited as an autosomal dominant trait, usually with reduced penetrance, but can also be inherited as a recessive trait. Digenic inheritance has also been reported as a likely inheritance pattern in an Australian family of European descent.[71] In this family, the inheritance pattern was consistent with autosomal dominance, however, two suggestive linkage regions with equal LOD scores were identified and all affected family members carried both disease-associated haplotypes. A two-locus linkage analysis was performed and together, the two linkage regions surpassed the significance threshold. In total, more than 20 significant or highly suggestive linkage regions for keratoconus have been identified to date.[28, 29, 64-66, 69, 71, 72, 74-80, 120, 141, 142] These linkage regions map to the majority of the autosomes and few regions have been replicated, suggesting that the linkage regions are largely family-specific. The only significant linkage region that has been replicated is a region on chromosome 5. The region 5q14.1-q21.3 reached significance in a four generation Caucasian family[74] and a suggestive peak was identified at 5q21.2 in a study of 23 small families from southern Italy.[141] An adjacent region on 5q also showed suggestive linkage with keratoconus in a study that investigated sibling pairs from a mixed population[75] and this region appears to overlap a highly suggestive linkage region at 5q31.1-q35.3 identified in an Ecuadorian family with autosomal dominant disease.[142] The

heterogeneity of keratoconus, along with the large number of genes harboured within the identified linkage regions, has made the identification of causal variants and key genes difficult.

Current candidate genes for isolated keratoconus identified through linkage studies include *IL1RN*,[66] *SLC4A11*[66] and *DOCK9*,[29, 120] however, as very little is known about the mechanism of disease, or key biological pathways involved, prioritising genes in linkage regions remains challenging. With the recent emergence of massively parallel sequencing, however, candidate genes within regions of interest can now be screened efficiently for variants without *a priori* hypotheses. It is therefore hypothesised that these technologies, paired with disease mapping in families with multiple cases of keratoconus and clear patterns of inheritance, will aid the identification of causative variants and key genes involved in keratoconus development and pathogenesis. A recent study demonstrates the potential of this strategy with the identification of a variant within the seed region of a non-coding RNA gene, *mir184*, via targeted sequencing of the genes within a linkage region on chromosome 15 in a family from Northern Ireland.[65, 77, 143] Affected individuals in this family had both keratoconus and congenital cataract and the same variant was found to co-segregate with keratoconus and congenital cataract in an unrelated family from Spain.[69] While this gene is involved a broader ocular phenotype, the identification of *mir184* in in these families provide a critical insight into biologically relevant pathways and may aid the identification of candidate genes in isolated keratoconus. The present study will go one step further by using whole genome sequencing (WGS) data to conduct disease gene mapping. Variation within any regions of interest, including both protein-coding and non-coding variants, will then be interrogated directly from the WGS data.

4.2 HYPOTHESIS AND AIM

The overarching hypothesis for this study was that rare, highly penetrant protein-coding variants contribute to keratoconus development. More specifically, we hypothesised that families with multiple cases of keratoconus with early onset or severe disease, and strong Mendelian inheritance patterns of disease, would aid the identification of putatively disease-causing variants involved in keratoconus development. This led to the following aim:

To identify rare, putatively disease-causing variants in families with multiple cases of early-onset or severe keratoconus and a strong Mendelian inheritance pattern of disease.

4.3 OVERALL STUDY DESIGN

This chapter focuses on the identification of putative disease-causing variants in two families: one with autosomal recessive inheritance of keratoconus and known consanguinity and another with apparent

autosomal dominant inheritance with reduced penetrance. This achieve this, WGS data was generated and used to conduct keratoconus mapping with an LD-pruned scaffold of SNPs that mimic a SNP array, as well as, investigate variation within the identified regions of interest. The WGS data were also used to determine the individuals' ancestry and confirm reported relationships in the families. A flow diagram of the study design is presented in Figure 4.1.

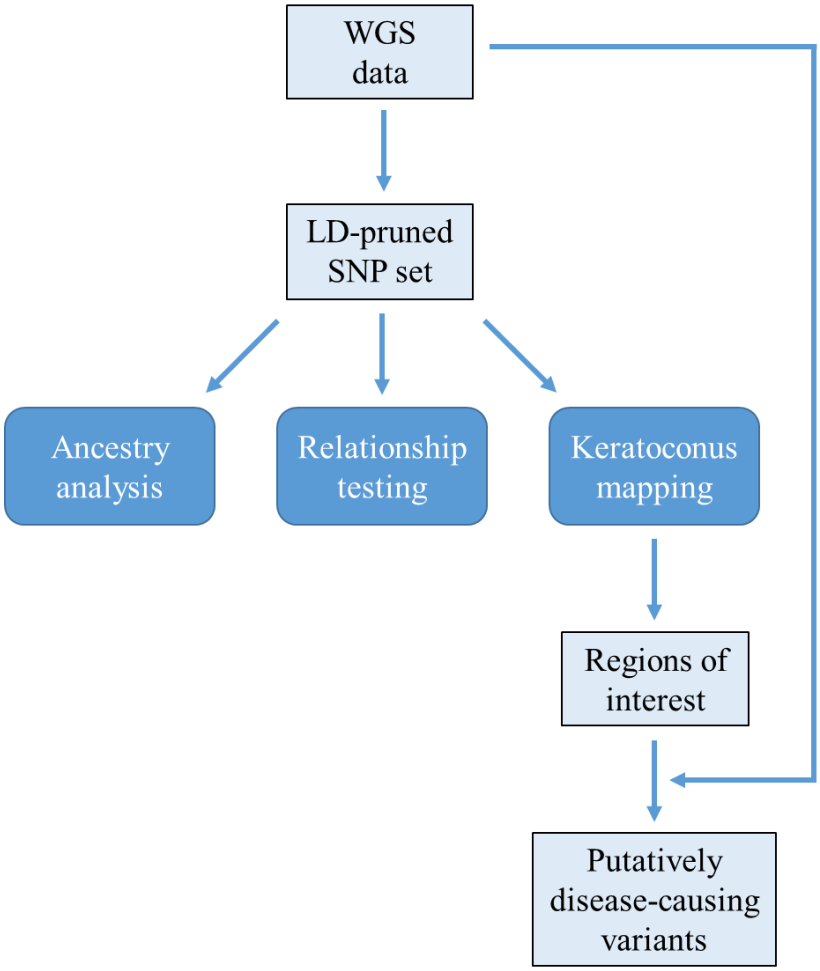


Figure 4.1 – A flow diagram of the overall study design.
Light blue boxes indicate data and dark blue boxes indicate methods. WGS = whole genome sequencing; LD = linkage disequilibrium; and SNP = single nucleotide polymorphism.

4.4 METHODS

4.4.1 Study participants

Families with multiple cases of keratoconus were recruited as described in Section 2.1.1.

KSA197

KSA197 is a small family of Italian heritage living in South Australia (Figure 4.2). The parents are second cousins and have two sons with keratoconus. Neither parent has keratoconus, but both have thin central corneas. The proband (KSA197.0) has a child (KSA197.4) that is reported as unaffected, however, has not been assessed by our ophthalmologists. A summary of the available clinical data is present in Table 4.1. The pattern of inheritance of keratoconus in this family is consistent with autosomal recessive disease.

DNA samples were collected from five family members: the two affected brothers, their parents and the unaffected child. WGS data was obtained for these five family members.

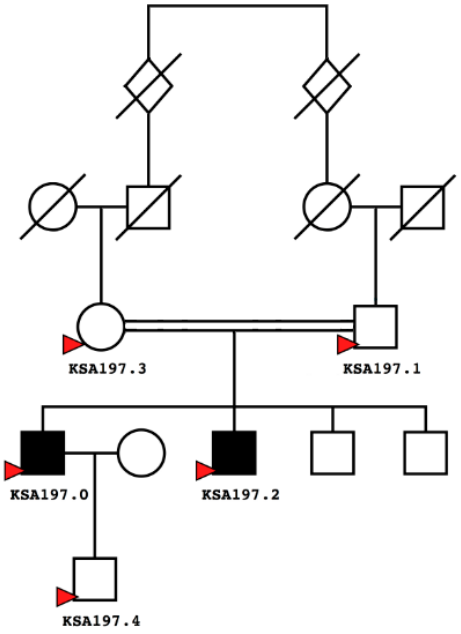


Figure 4.2 – The KSA197 family pedigree.
Females are indicated by circles, males by squares, individuals with unknown sex by diamonds, and deceased individuals are depicted by a slash. These symbols are coloured black for individuals with keratoconus and white for unaffected individuals. Red arrows indicate that DNA was collected and whole genome sequencing (WGS) data was generated.

Table 4.1 – Clinical data for KSA197 family members.

Individual	Keratoconus Status	K readings		CCT (µm)	
		RE	LE	RE	LE
KSA197.0	affected	40.25/40.5	52.25/54.00	497	385
KSA197.1	unaffected	43.0/41.87	42.75/42.37	509	521
KSA197.2	affected	46.75/48.20	44.60/45.00	378	408
KSA197.3	unaffected	42.75/42.75	43.0/42.75	480	473
KSA197.4	unaffected*	NA	NA	NA	NA

K = keratometry; a measure of the curvature of the cornea in dioptres (D).
CCT = central corneal thickness; a measure of the thickness of the cornea.
RE = right eye.
LE = left eye.
NA = not assessed.
* Self reported.

KCNSW01

KCNSW01 is a Jordanian family with eight individuals with severe and early-onset keratoconus (Figure 4.3). The proband, KCNSW01-2, is a severely affected child who was diagnosed at five years of age. KCNSW01-2 is known to have a *de novo* deletion at 15q26.3 (chr15:96,725,466-100,200,967), but no other clinical details were disclosed. The proband and his father (KCNSW01-1), who is mildly affected, live in Sydney, where they were recruited for this study. A severely affected uncle (KCNSW01-3) also lives in Sydney. All remaining family members live in Jordan, however written informed consent was obtained, and samples were collected remotely by saliva collection kit. The proband's father reports severe and early-onset disease (development at 6-7 years) in three of his four sisters, and variable severity in three of his six brothers. The parents of these individuals are reported to be unaffected and unrelated, although these individuals have not been formally examined for mild disease. In total, seven of the 11 siblings in the second generation of KCNSW01 are affected with keratoconus.

The second and third generation of KCNSW01 is consistent with autosomal dominant inheritance, however, the parents in the first generation are reported as unaffected. Both parents have 9 siblings, all of whom are reported to be unaffected with keratoconus. KCNSW01-1 reported that there was no consanguinity in recent generations. Saliva samples for DNA extraction were collected from eleven individuals across the three generations.

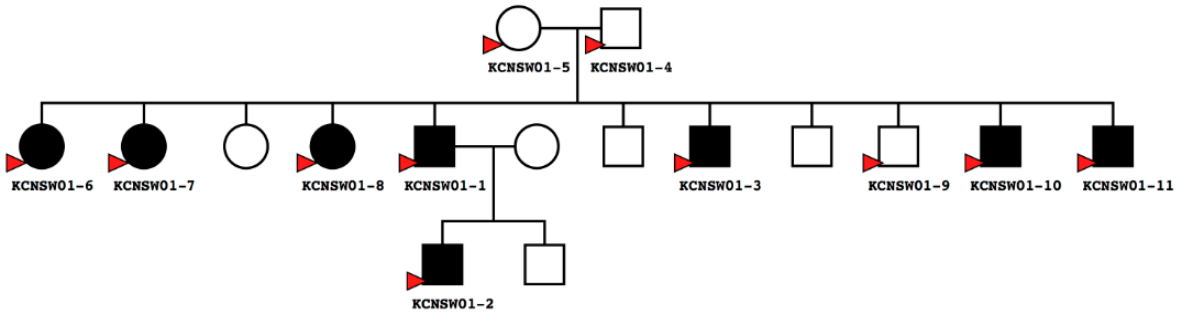


Figure 4.3 – The KCNSW01 family pedigree.
Females are indicated by circles and males by squares. These symbols are coloured black for individuals with keratoconus and white for unaffected individuals. Red arrows indicate that DNA was collected and WGS data was generated.

4.4.2 Whole genome sequencing

Whole genome sequencing was performed using Illumina's TruSeq Nano Library Prep v2.5 with 150 bp paired-end sequencing on the Illumina XTen platform (30x coverage) at the Kinghorn Centre for Clinical Genomics (Sydney, Australia). All family members across KSA197 and KCNSW01 who provided DNA samples were sequenced (n=16). Raw sequence data were provided as unmapped FASTQ files. These data were mapped to the human reference genome (hg19) using BWA[188] and variants were joint-called using GATK[125] in concordance with GATK's "Best Practices" guidelines in-

house using the Churchill[162] pipeline. This resulted in variant caller format (VCF) files for SNPs and small insertions and deletions (indels).

4.4.3 Generating a linkage disequilibrium-pruned SNP set for LD-sensitive analyses

A linkage disequilibrium (LD)-pruned SNP set was generated using an independent cohort consisting of 1585 unrelated Caucasians genotyped on the Human Omni Express v1.1 BeadChip (Illumina). The cohort has previously been described in detail.[189] Using these data, pairwise LD SNP pruning was performed for autosomal SNPs in PLINK[190] (version 1.90) with the following parameters: window size=50, step=5, pairwise r^2 threshold=0.5. This identified a SNP set of 266,025 LD-pruned SNPs. For each family separately, the genotypes at these SNPs were extracted from the joint-called VCF files using version 1.5 of BCFtools (available from https://github.com/samtools/BCFtools). An example command is outlined in Appendix 4. If the alternate allele was not observed in any of the family members, the SNP was not captured in the family-specific VCF file. A total of 191,684 SNPs were included for KCNSW01 and 189,747 SNPs were included for KSA197.

To ensure only high confidence genotypes were included in subsequent analyses, confidence tags were added to the genotypes in the VCF as described in Section 2.2.1. 'Clean' VCF files were generated by converting any genotypes with low coverage or low quality scores to missing calls ('./.') as outlined in Section 2.2.2. For the KSA197 family, this 'clean' VCF file was used for homozygosity mapping as described in Section 4.4.6.1. Additionally, PLINK binary files were created by reading the 'clean' VCF files into PLINK using the '--vcf' and '--keep-allele-order' parameters and the PLINK format sample information (FAM) files were manually updated with the family IDs, parent IDs, and sex and phenotype information. The PLINK binary files were subsequently used for LD-sensitive analyses such as determining the ancestry of the families (Section 4.4.4), relationship testing (Section 4.4.5), parametric linkage analysis for KCNSW01 (Section 4.4.7.1), identifying runs of homozygosity in KSA197 using PLINK (Section 4.4.6.1), and haplotype analysis for both families (Sections 4.4.6.2 and 4.4.7.2).

4.4.4 Determining ancestry using principle components analysis

Principle components analysis (PCA) was used to determine the ancestry of KCNSW01 and KSA197 on a continental-scale using the International HapMap[191] Project Phase III (HapMap3) genotype data from Europe (CEU), Asia (CHB + JPT) and Africa (YRI) as described by Anderson and colleagues.[192] Using PLINK, a list of SNPs that were common between the LD-pruned SNP-set in the two families (described in Section 4.4.3), as well as, the HapMap3 data for the four ethnic populations was generated. Multi-allelic SNPs and A/T and C/G SNPs were excluded. 145,229 SNPs were common to the genotype data in both families and the HapMap3 data. Using PLINK, the genotype data for these SNPs from the families and HapMap3 were merged and PCA was performed using EIGENSTRAT.[193] A scatter plot

of the first two principle components was generated using the ggplot2[172] package in R.[126] An example of the code presented in Appendix 5.

4.4.5 Relationship testing

To confirm the reported relationships in each family, identity-by-descent (IBD) estimation was conducted in PLINK, using the '--genome parameter'. The family-specific genotype data for the LD-pruned SNP set (as described in section 4.4.3) was used for this analysis. For each pair of individuals included in the analysis, PLINK estimates the probability that zero alleles are IBD (Z0), the probability that one allele is IBD (Z1), and the probability that two alleles are IBD (Z2) at a given locus. PLINK also estimates the summary statistic PI_HAT which represents the overall IBD proportion.

To visualise the output, 3D plots (Z0 vs Z1 vs Z2) were generated using the Plotly[194] package in R. To easily distinguish between the relationship clusters, pairs of individuals were coloured based on their reported relationship. Example code is available in Appendix 6.

4.4.6 Keratoconus mapping in KSA197

4.4.6.1 Homozygosity mapping

Due to the known relationship between the parents and autosomal recessive inheritance pattern of keratoconus in KSA197, it was hypothesised that the causative variant was located in a homozygous region shared by the two affected brothers. Homozygosity mapping was conducted using the online program HomozygosityMapper[195] (available at http://www.homozygositymapper.org/) using the 'clean' VCF file containing only high confidence variant calls for the 189,747 LD-pruned SNPs present in the family (described in 4.4.3). Homozygosity was required in both cases (KSA197.0 and KSA197.2), however, homozygous stretches longer than 2,000 bp that were also homozygous in the unaffected family members (KSA197.1, KSA197.3 and KSA197.4) were excluded. The default maximum block length of 250 SNPs was applied. Homozygosity scores were downloaded and plotted using the ggplot2 package in R as outlined in Appendix 7. In addition, runs of homozygosity were identified in PLINK using the binary PLINK files generated in Section 4.4.3. An example command is outlined in Appendix 8. This analysis identifies runs of homozygosity shared between pairs of individuals, however, only regions shared by the two affected brothers (KSA197.0 and KSA197.2) were of interest.

4.4.6.2 Identifying regions of interest

Regions identified by both HomozygosityMapper and PLINK's runs of homozygosity utility were selected as regions of interest in KSA197 and further investigated. Haplotype estimation was used to confirm the homozygosity region and determine the disease-associated haplotype.

The binary PLINK files containing high confidence genotypes for the family members at the LD-pruned SNP set described in Section 4.4.3 were used for this analysis. The centimorgan (cM) positions for the SNPs were extracted from SHAPEIT-format recombination map files for the HapMap phase II b37 data (available from: http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#gmap) using the '--cm-map' flag in PLINK. Mega2[196] was used to convert these PLINK files to MERLIN[197] input format files. A custom R script, developed by Ms. Johanna Jones (Menzies Institute for Medical Research, University of Tasmania, TAS, Australia) and Mr. Michael Sumner (University of Tasmania, TAS, Australia), was applied to the Merlin format map files to identify and amend non-unique cM positions caused by the large number of included variants (Appendix 9). This script identified pairs of SNPs with duplicate cM position and adds a very small number (0.0000001) to the cM position of one SNP in each pair, resulting in unique cM positions without affecting the linkage analysis results. Unlikely genotypes were detected using the '--error' flag in MERLIN[197] and were removed from analysis using MERLIN's utility program Pedwipe.

Haplotype estimates for the relevant chromosome was subsequently generated using the most likely haplotype reconstruction in MERLIN. The regions of interest were manually inspected to determine the disease-associated haplotype. An example of the MERLIN command is outlined in Appendix 10.

4.4.6.3 Extracting variants from the whole genome sequencing data

SNPs and small indels called in the KSA197 family members within the region of interest were extracted from WGS data using BCFtools (available from https://github.com/samtools/BCFtools). These variants were annotated as described in Section 2.2.3. Key annotations included variant identification codes (IDs) from the dbSNP[128] 147 database; genes from the RefGene[128] database; minor allele frequencies (MAF) from the Genome Aggregation Database[129] (gnomAD); and deleteriousness/pathogenicity predictions for SNPs from the Combined Annotation Dependent Depletion[134] (CADD) and Functional Analysis through Hidden Markov Models[135] (FATHMM), specifically the FATHMM-MKL[136] algorithm for SNPs and the FATHMM-indel[137] algorithm for small insertions or deletions (indels).

Coverage in the WGS data across the region of interest in all family members was determined using the depth utility in SAMTools[164] (version 1.8). Using R, the mean depth and the standard deviation was calculated across all bases within the region and all family members. Bases with a mean depth < 10 were written to a BED format file in R and, using the merge utility of BEDTools[198] (version 2.26.0), consecutive bases were combined into a single region. Examples of these commands are outlined in Appendix 11.

4.4.6.4 Identifying putatively disease-causing variants

As the inheritance pattern of keratoconus in KSA197 is consistent with recessive disease and there is a history of consanguinity, it was hypothesised the causative variant is present in the homozygous state in the two affected brothers. Segregating variants were therefore classified as those that were homozygous in KSA197-0 and KSA197-2, but heterozygous in the unaffected parents and child (KSA197-1, KSA197-3 and KSA197-4), within the region of interest. It was also hypothesised that the causative variant was likely to be rare in the general population and have an intermediate prediction of deleteriousness/pathogenicity, as it only causes disease in the homozygous state. Based on these hypotheses, two filtering strategies were designed to identify putatively disease-causing variants in KSA197.

Filtering Strategy 1

To address the overarching hypothesis of the study, Filtering Strategy 1 included rare protein-coding variants located within the homozygosity region. More specifically, putatively-disease causing variants were defined as those that were located in protein-coding regions, were rare (MAF < 1%) in all populations in gnomAD, segregated, and had either a CADD score ≥ 10 or a FATHMM score ≥ 0.5.

Filtering Strategy 2

As WGS data was available, Filtering Strategy 2 did not limit the analysis to protein-coding variants and classified putatively disease-causing variants as those that segregated, were rare (MAF < 1%) in all populations in gnomAD, and either had a CADD score ≥ 10 or a FATHMM score ≥ 0.5.

4.4.7 Keratoconus mapping in KCNSW01

4.4.7.1 Parametric linkage analysis

Parametric linkage analysis was performed for KCNSW01 as the pattern of keratoconus in this family was consistent with autosomal dominant inheritance with reduced penetrance. MERLIN format files were generated for KCNSW01 using the method described in Section 4.4.6.2. To complete the pedigree, the mother of KCNSW01-2 was manually added to the PLINK files prior to the conversion of PLINK files to MERLIN files, however as this individual wasn't sequenced all genotypes were coded as missing. As the parents in the first generation (KCNSW01-4 and KCNSW01-5) are reported as unaffected individuals, but have not been examined by our ophthalmologists, they were coded with an unknown phenotype.

As the second and third generation of KCNSW01 was consistent with an autosomal dominant inheritance of keratoconus, parametric linkage analysis was performed in MERLIN using a dominant model with reduced penetrance (0.0001, 0.9, 1). The detection and removal of unlikely genotypes, as well as the parametric linkage analysis for all autosomes, was automated using a custom script presented in Appendix 12. The tabulated output files for each autosome were concatenated and an autosome-wide

plot was generated using the ggplot2 package in R. The code used to generate the plot is outlined in Appendix 13.

4.4.7.2 Identifying regions of interest

Regions of interest were defined as regions with a parametric LOD score above 2. To determine the disease-associated haplotypes and assess the coverage across the regions of interest, haplotype analysis was performed in MERLIN as outlined in Section 4.4.6.2.

4.4.7.3 Identifying putatively disease-causing variants

SNPs and small indels called in the KCNSW01 family members within the regions of interest were extracted from the WGS data using BCFtools and were annotated as previously described in Section 4.4.6.3 (for KSA197). Segregating variants were classified as heterozygous variants that were present in individuals all carrying the disease-associated haplotype and absent in non-carriers, regardless of affection status. As two regions of interest were identified in KCNSW01, two hypotheses were developed to aid the identification of putatively disease-causing variants.

The first hypothesis states that one of these regions is real and harbours the causative variant, while the other linked region co-segregates with disease by chance. This hypothesis is consistent with an autosomal dominant inheritance pattern of disease and from here on will be called the 'autosomal dominant hypothesis'. Under this hypothesis, putatively disease-causing variants are likely to be rare and predicted to be highly deleterious/pathogenic, however two distinct filtering strategies were used to investigate the role of both protein-coding variation and non-coding variants. Filtering Strategy 1 was designed to address the overarching hypothesis of the study and therefore putatively disease-causing variants were defined as variants that were located within protein-coding regions, segregated with the disease-associated haplotypes, were rare (MAF < 0.01) in all populations in gnomAD, and had either a scaled CADD score ≥ 15 or a FATHMM-MKL/FATHMM-indel score ≥ 0.8. Filtering Strategy 2 was very similar but did not limit the analysis to protein-coding variants. Under Filtering Strategy 2, putatively disease-causing variants included those that segregated with the disease-associated haplotypes, were rare (MAF < 0.01) in all populations in the gnomAD database and had either a scaled CADD score ≥ 15 or a FATHMM-MKL/FATHMM-indel score ≥ 0.8.

The second hypothesis, the 'digenic hypothesis' states that keratoconus in KCNSW01 is inherited in a digenic fashion and therefore one variant from both of the disease-linked regions is required, and only in combination, cause keratoconus. Qiagen's Ingenuity Pathways Analysis (IPA) was used to identify published interactions between pairs of genes located within the risk-associated haplotypes. Gene sets containing the list of genes located within the regions of interest were added to the 'my pathways' tool and the Ingenuity Knowledge Base was interrogated to identify direct and indirect interactions between pairs of gene products, including protein-protein binding interactions and gene expression changes.

Under the digenic hypothesis, putatively disease-causing variants were not required to both be rare, but in combination must be rare, therefore no MAF filter was applied under this hypothesis as even the most common variant observed together with a novel variant is a rare event. Furthermore, as it is hypothesised that neither variant is sufficient to cause disease in isolation, a lower threshold of deleteriousness/pathogenicity was considered. Therefore, putatively disease-causing variants were defined as those that were located within potentially interacting genes (or located the flanking non-coding regions) with a scaled CADD score ≥ 10 and/or a FATHMM-MKL/FATHMM-indel score ≥ 0.5.

4.4.8 Interrogating putatively disease-causing variants identified in the families

Putatively disease-causing variants were further prioritised manually as described in Section 2.2.4.

4.5 RESULTS

4.5.1 Determining ancestry using principle components analysis

The PCA analysis demonstrated both families were of largely European ancestry (Figure 4.4). The individuals cluster in family groups on the European-African axis as expected, with the KSA197 family members clustering slightly closer to the Northern and Western European population from HapMap3 than KCNSW01.

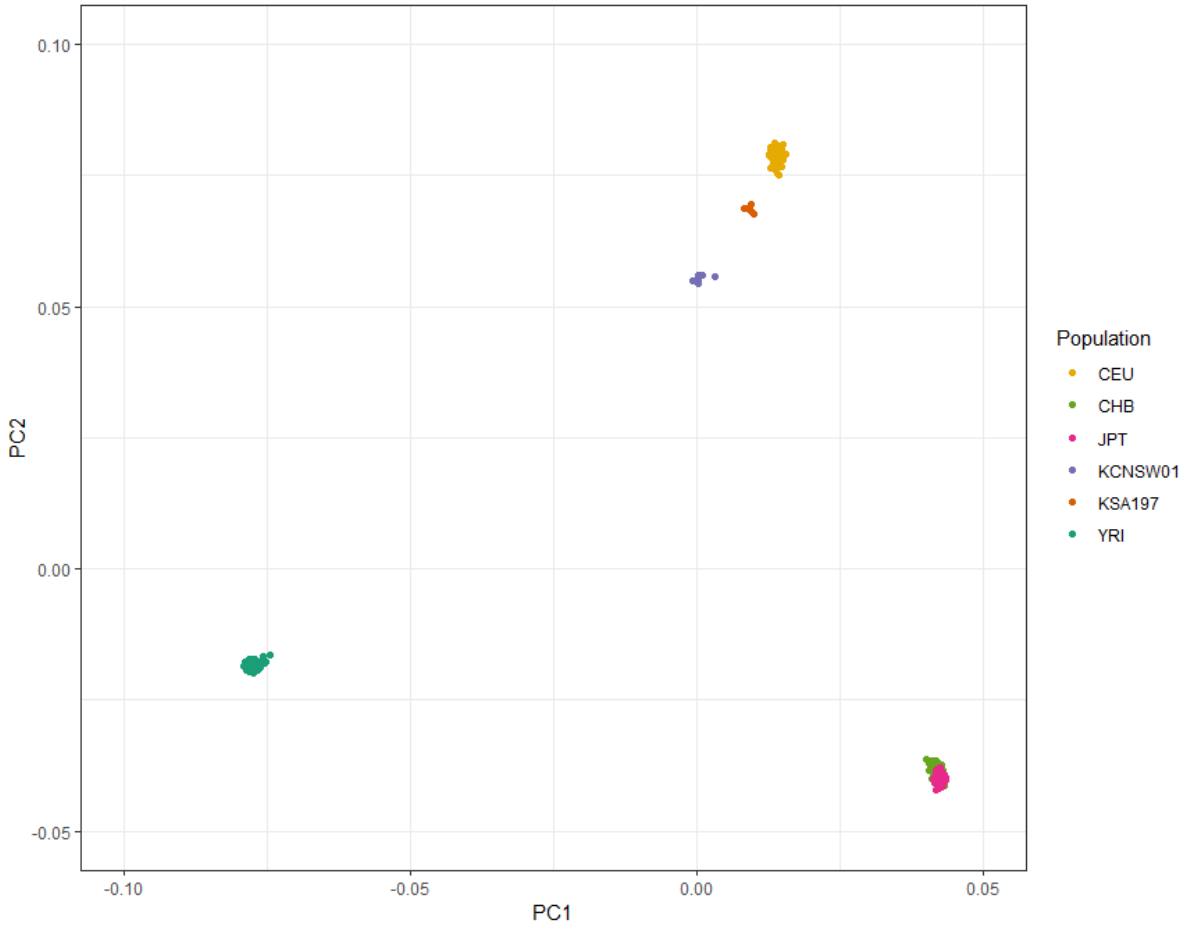


Figure 4.4 – A scatter plot for the first two principle components including individuals in HapMap Phase III and family members from KCNSW01 and KSA197.
CEU = Utah residents with Northern and Western European ancestry; CHB = Han Chinese living in Beijing, China; JPT = Japanese living in Tokyo, Japan; YRI = Yoruba living in Ibadan, Nigeria.

4.5.2 KSA197

4.5.2.1 Relationship testing

The recorded relationships for KSA197 were confirmed by IBD estimation (Figure 4.5). The full IBD estimates for each pair of individuals are outlined in Appendix 14. As expected, KSA197.1 and KSA197.3 were related (PI-HAT = 0.40), however, all PI-HAT estimates were more inflated than expected, indicating that a family history of consanguineous unions is likely.

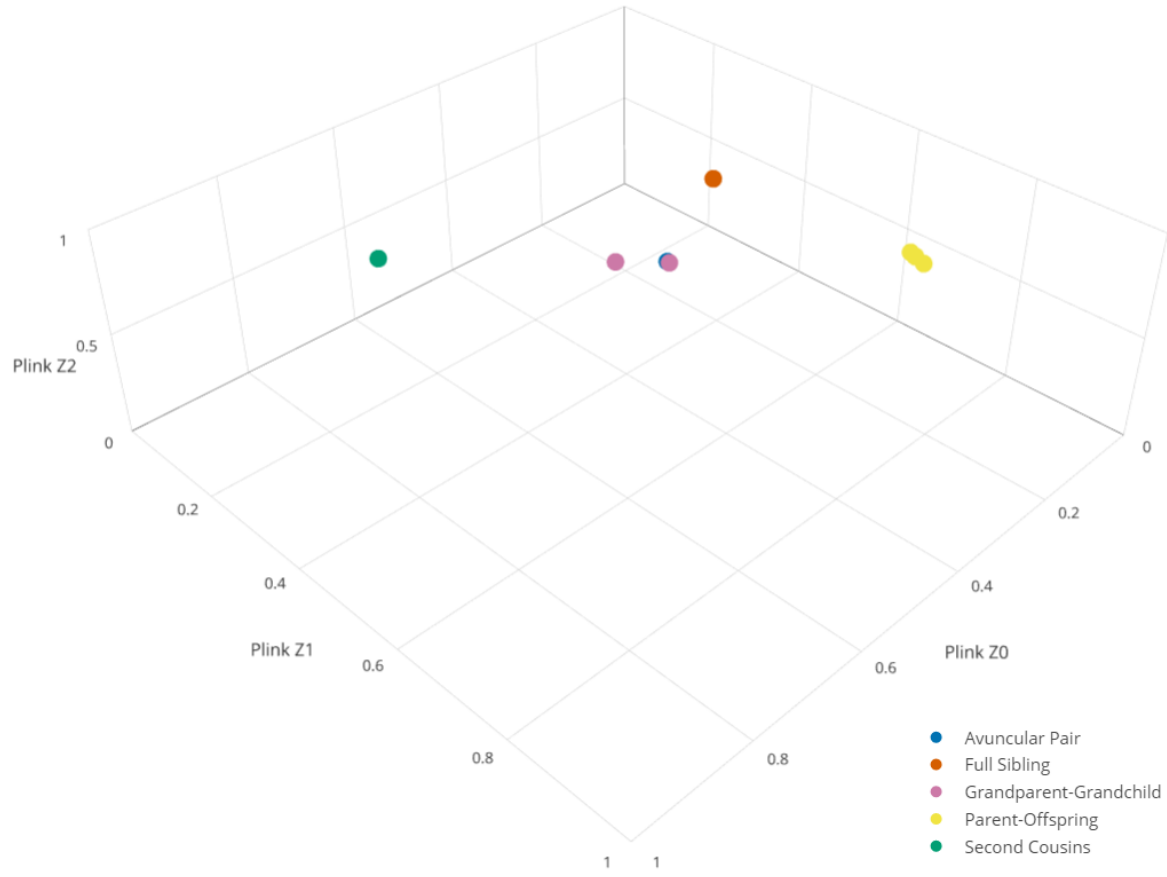


Figure 4.5 – A 3D plot of the PLINK identity-by-descent estimates for the KSA197 family members.

The coloured data points indicate the relationship between a pair of individuals where green= the parents (second cousins), blue = avuncular pairs, pink = grandparent-grandchild pairs, orange = full sibling pairs, and yellow = parent-offspring pairs. This plot was generated using the Plotly[194] package in R.[126] An online and interactive version of this plot is available at **https://plot.ly/~semlucas/7**.

4.5.2.2 Homozygosity mapping

A region on chromosome 16 was identified by both HomozygosityMapper and the runs of homozygosity analysis in PLINK. A 1.2 Mb region between rs237135 and rs1363749 (spanning chr16:26642059-27827381) obtained the maximum homozygosity score of 342 using HomozygosityMapper (Figure 4.6). Similarly, the region between rs237135 and rs4787993 (chr16:26642059-27827858) was flagged by PLINK as a run of homozygosity shared by the two affected brothers (KSA197.0 and KSA197.2; Table 4.2). This was the only run of homozygosity identified by PLINK that was shared by the brothers. The region identified by PLINK encompasses the homozygous region identified by HomozygosityMapper, however, the run of homozygosity region extends an additional 477 bp downstream. Haplotype analysis in MERLIN confirmed the homozygous region shared by KSA197.0 and KSA197.2 extended from rs237135 to rs1363749 (chr16:26642059-27827381) and this region was selected as a region of interest for further investigation.

Table 4.2 – The run of homozygosity shared by KSA197.0 and KSA197.2

Region	Upstream flanking SNP	Downstream flanking SNP	Region Size (kb)	Number of SNPs in run
chr16:26642059–27827858	rs237135	rs4787993	1185.8	174

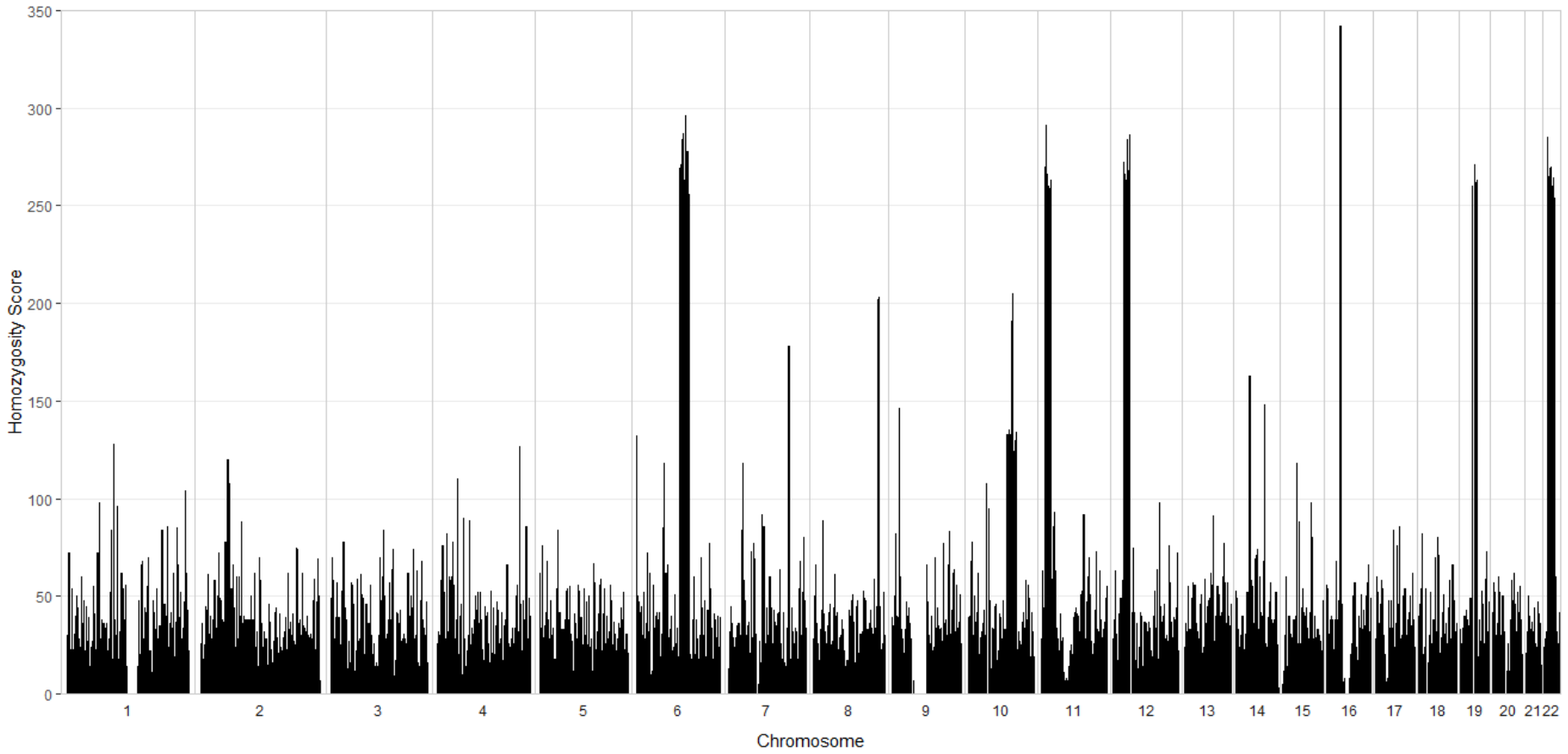


Figure 4.6 – Autosome-wide homozygosity scores for KSA197.
Homozygosity scores were obtained from HomozygosityMapper[195] and plotted using the ggplot2[172] package in R.[126]

4.5.2.3 Coverage across the homozygous region

The homozygous region (chr16:26642059-27827381) had mean depth of 38.4 reads (standard deviation, sd = 5.5) across the five KSA197 family members. Only 2,263 bases within this region had a mean read depth below 10 (0.19% of bases). These regions are likely to have insufficient coverage for high confidence variant calls in the family members, may or indicate deletions in multiple family members.

4.5.2.4 Identifying putatively disease-causing variants

The homozygosity region harboured 204 SNPs, 17 insertions and 29 deletions that segregated with keratoconus in KSA197 (Table 4.3). It's important to note that 306 indels were called in the heterozygous state in the two affected brothers, despite being located within the homozygosity region. Upon close inspection, however, all of these heterozygous calls were located at repetitive regions and were largely of poor confidence with a read depth below 10 and/or a quality score below 20.

Filtering Strategy 1

No variants fulfilled the criteria for Filtering Strategy 1. In fact, only one exonic variant segregated with disease within the homozygosity region: a nonsynonymous variant in *KIAA0556* (Table 4.4). This variant, however, has a MAF ≥ 1.3% in the non-Finnish European, admixed American and 'Other' populations in gnomAD. Furthermore, the variant has been observed in the homozygous state in 43 individuals in the non-Finnish European population in gnomAD corresponding to a frequency of 6.8 $x10^{-4}$ which is roughly equal to the overall prevalence of keratoconus (1 in 1,500) in Europeans. As keratoconus is a highly heterogeneous disease, and the causative variant in KSA197 may be private to this family or contribute to a very small proportion of keratoconus cases, this variant is too common to account for disease in this family.

Filtering Strategy 2

Ten variants were rare, including 5 novel variants, and these variants are presented in Table 4.5. Of the ten rare variants, only one fulfilled the criteria for putatively disease-causing under Filtering Strategy 2: a novel SNP between *C16orf82* and *KDM8* (chr16:27115706C>A). This variant obtained a CADD score of 22, a FATHMM score of 0.96 and occurs within a DNaseI hypersensitive region identified in osteoblasts. While the alternate allele (A) identified in KSA197 is not observed in other vertebrate species, the reference allele is a 'T' in dog, wallaby and green sea turtle and therefore this site is not completely conserved across vertebrates.

Table 4.3 – A summary of the segregating variants identified in KSA197 within the chromosome 16 homozygous region.

	Variant Type				Variant Location				Highly Prioritised Variants				
Region	**SNPs**	**Ins.**	**Del.**	**Total**	**Exonic (NC)**	**Intronic (NC)**	**UTRs**	**Intergenic**	**Novel**	**Rare**	**CADD ≥ 10 || FATHMM ≥ 0.5**	**FS1**	**FS2**
chr16:26642059-27827381	204	17	29	250	1 (0)	21 (1)	1	227	5	10	24	0	1

SNPs = the number of single nucleotide polymorphisms.
Ins. = the number of insertions.
Del. = the number of deletions.
NC = the number of non-coding RNA variants included in the variant counts for exonic and intronic variants.
UTRs = the number of variants located in untranslated regions (either 3' or 5').
Novel = the number of variants without minor allele frequencies in any of the populations in the gnomAD database, Kaviar or 1KGP (August 2015 release).
Rare = the number of variants with a minor allele frequency (MAF) less than 1% in all populations of the gnomAD database.
CADD ≥ 10 || FATHMM ≥ 0.5 = the number of variants with a scaled CADD score of at least 10 and/or a FATHMM-MKL/FATHMM-indel score of at least 0.5.
FS1 = the number of variants that fulfilled the criteria for putatively disease-causing variants under Filtering Strategy 1 (rare protein-coding).
FS2 = the number of variants that fulfilled the criteria for putatively disease-causing variants under Filtering Strategy 2 (not limited to protein-coding).

Table 4.4 – The segregating exonic variant located within the homozygosity region in KSA197.

Position	Ref	Alt	SNP ID*	Gene	Variant	gnomAD max	gnomAD NFE	CADD	FATHMM
chr16:27784497	G	A	rs117316062	*KIAA0556*	c.G4276A; p.(E1426K)	0.0245	0.0245	24.60	0.96

Ref = reference allele.
Alt = alternate allele.
Variant = the nucleotide variant and inferred protein variant are presented.
gnomAD max = the maximum alternate allele frequency observed across all populations in the gnomAD database.
gnomAD NFE = the non-Finnish European population of the gnomAD database.
CADD = the scaled CADD score.
FATHMM = FATHMM score for coding variants.

Table 4.5 – Rare, segregating variants identified in KSA197 within the chromosome 16 homozygosity region.

Position	Ref	Alt	SNP ID	Gene	gnomAD max	gnomAD NFE	CADD	FATHMM
chr16:26760228	G	A	rs758556074	*HS3ST4 – C16orf82*	0.0004	$4.00x10^{-4}$	5.95	0.1159
chr16:26777357	-	AGATGATAGAT	novel	*HS3ST4 – C16orf82*	0.0000	0.0000	0.22	NA
chr16:26867292	G	T	novel	*HS3ST4 – C16orf82*	0.0000	0.0000	0.10	0.04964
chr16:26934160	T	C	novel	*HS3ST4 – C16orf82*	0.0000	0.0000	1.97	0.07106
chr16:26958687	A	G	rs141230258	*HS3ST4 – C16orf82*	0.0099	0.0052	3.14	0.09619
chr16:27030292	C	T	rs183043552	*HS3ST4 – C16orf82*	0.0010	0.0000	1.48	0.06647
chr16:27115706	**C**	**A**	**novel**	***C16orf82 – KDM8***	**0.0000**	**0.0000**	**22.00**	**0.96131**
chr16:27119478	G	A	novel	*C16orf82 – KDM8*	0.0000	0.0000	5.39	0.0805
chr16:27432490	T	G	NA	*IL21R* (intronic)	0.0003	$3.00x10^{-4}$	6.87	0.23092
chr16:27659545	T	G	NA	*KIAA0556* (intronic)	0.0001	$1.00x10^{-4}$	3.14	0.06869

Position = position refers to the position of SNPs and the start position for indels.
Ref = reference allele.
Alt = alternate allele.
gnomAD max = the maximum alternate allele frequency observed across all populations in the gnomAD database.
gnomAD NFE = the non-Finnish European population of the gnomAD database.
CADD = the scaled CADD score.
FATHMM = FATHMM score for non-coding variants.
NA = this variant wasn't attributed an ID in the dbsnp 147 database.
The variant that fulfilled the criteria for putatively disease-causing under Filtering Strategy 2 is **bold**.

4.5.3 KCNSW01

4.5.3.1 Relationship testing

The IBD estimation confirmed all reported relationships for KCNSW01 (Figure 4.7). The full IBD estimates for each pair of individuals are outlined in Appendix 14. As seen in the plot, avuncular pairs (uncle/aunt- nephew pairs) and grandparent-grandchild pairs clustered together as they are both second degree relationships, but all other relationship groups cluster separately. The parents obtained a PI-HAT of 0.19, thus demonstrating that they are related, but by more than 2 degrees of separation.

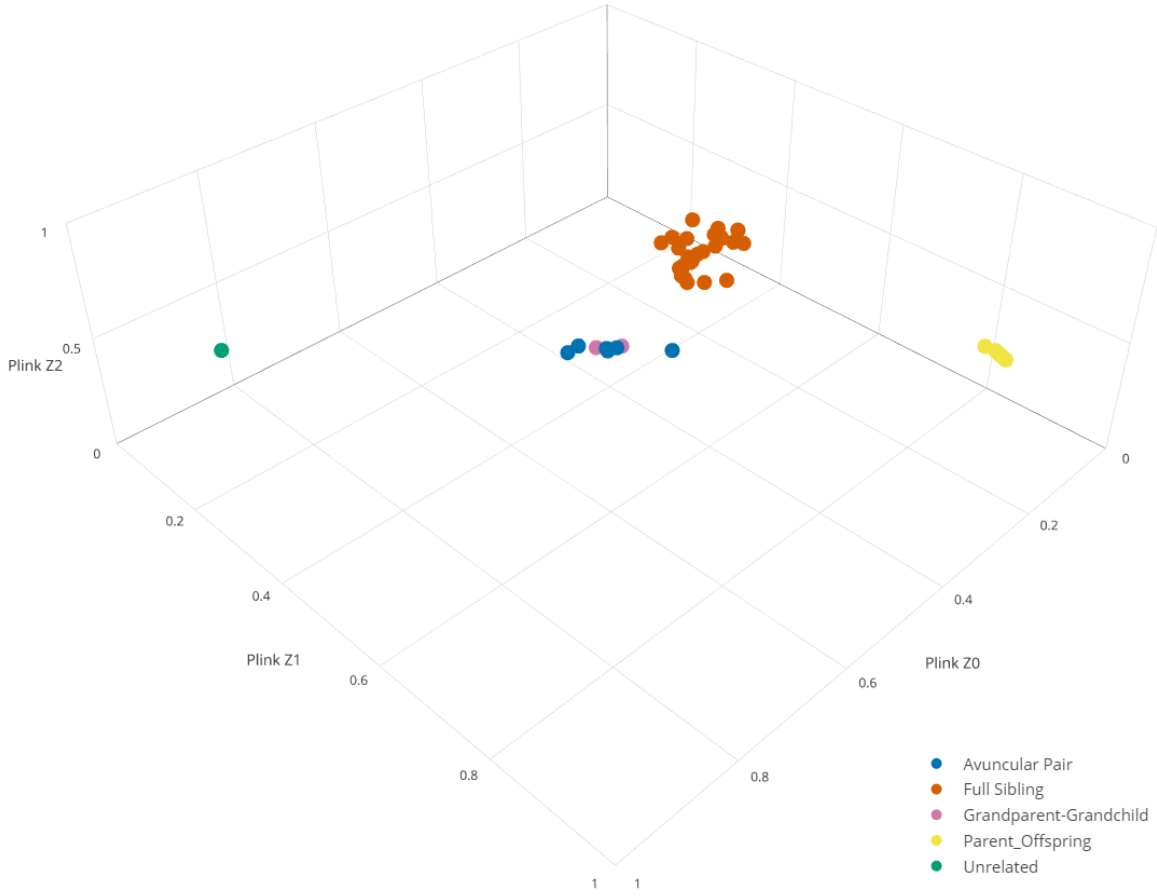


Figure 4.7 – A 3D plot of the PLINK identity-by-descent estimates for the KCNSW01 family members.
The coloured data points indicate the relationship between a pair of individuals where green= unrelated pairs (the parents), blue = avuncular pairs, pink = grandparent-grandchild pairs, orange = full sibling pairs, and yellow = parent-offspring pairs. This plot was generated using the Plotly[194] package in R.[126] An online and interactive version of this plot is available at **https://plot.ly/~semlucas/9**.

4.5.3.2 Parametric linkage analysis

Linkage analysis identified two regions with maximum LOD scores of 2.06: a 1.5 Mb region between rs2009472 and rs3213636 at 17q12 and a 10.2 Mb region flanked by rs6040904 and rs6041016 at

20p13-12.2 (Figure 4.8). Analysis of the haplotypes at these loci demonstrated that the risk-associated haplotype at 17q12 was inherited from the matriarch (KCNSW01-5), the risk-associated haplotype at 20p13-12.2 was inherited from the patriarch (KCNSW01-4), and all individuals with keratoconus have inherited both of these haplotypes (Figure 4.9). The only unaffected individual sequenced in the second generation, KCNSW01-9, did not inherit either of the risk-associated haplotypes. Based on these observations, segregating variants for the 17q12 locus were defined as those that were present in all affected individuals and KCNSW01-5, but absent in KCNSW01-9. Similarly, segregating variants at the 20p13-12.2 locus were variants that were observed in all affected individuals and KCNSW01-5, but absent in KCNSW01-9.

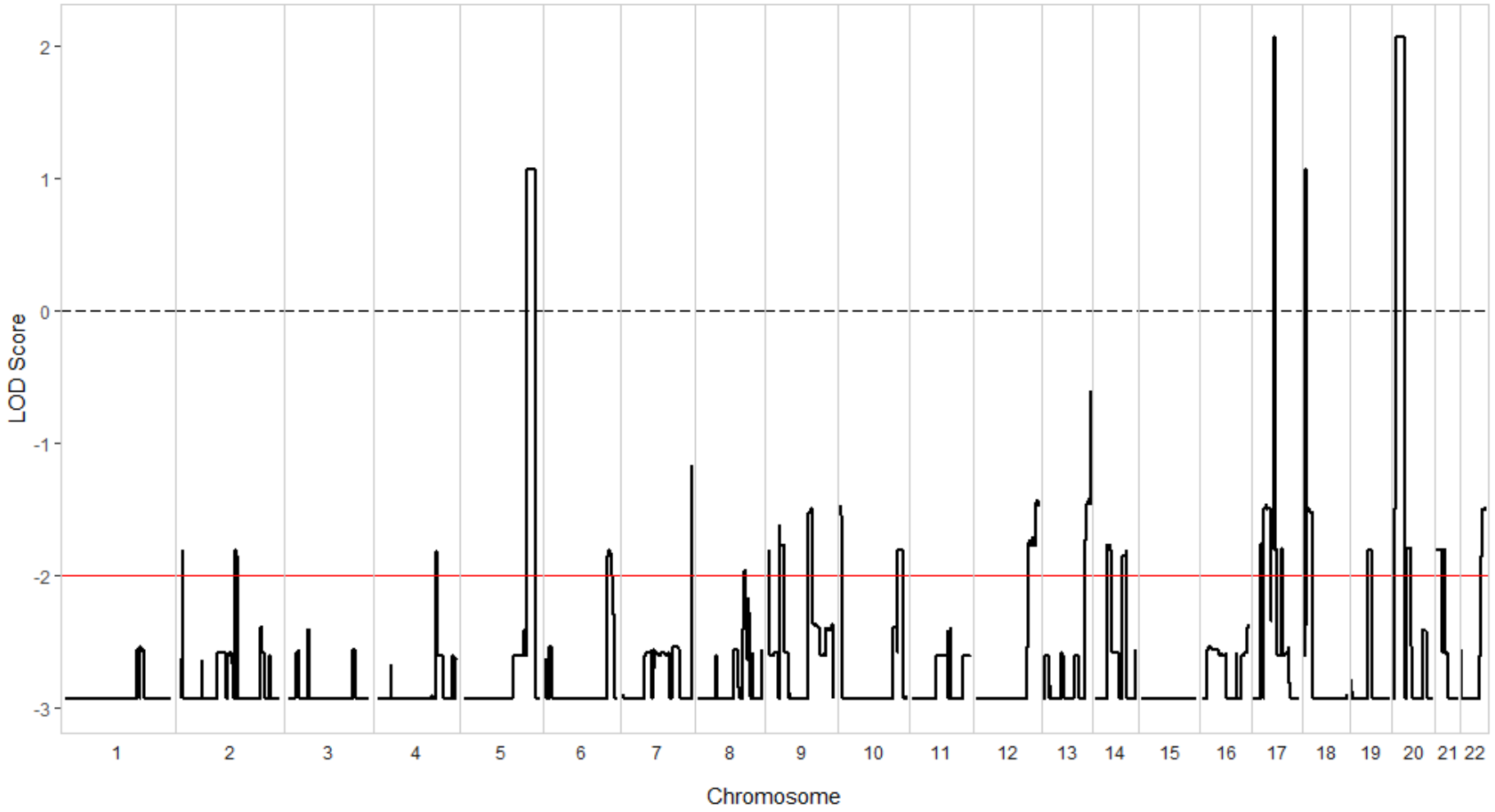


Figure 4.8 – Autosome-wide parametric linkage analysis results for KCNSW01.
The dashed line indicates a LOD score of 0 and the red line indicates a LOD score of -2. Any region that does not surpass a LOD score of -2 has significant evidence against linkage. LOD scores were generated using parametric linkage in MERLIN[197] and plotted with the ggplot2[172] package in R.[126]

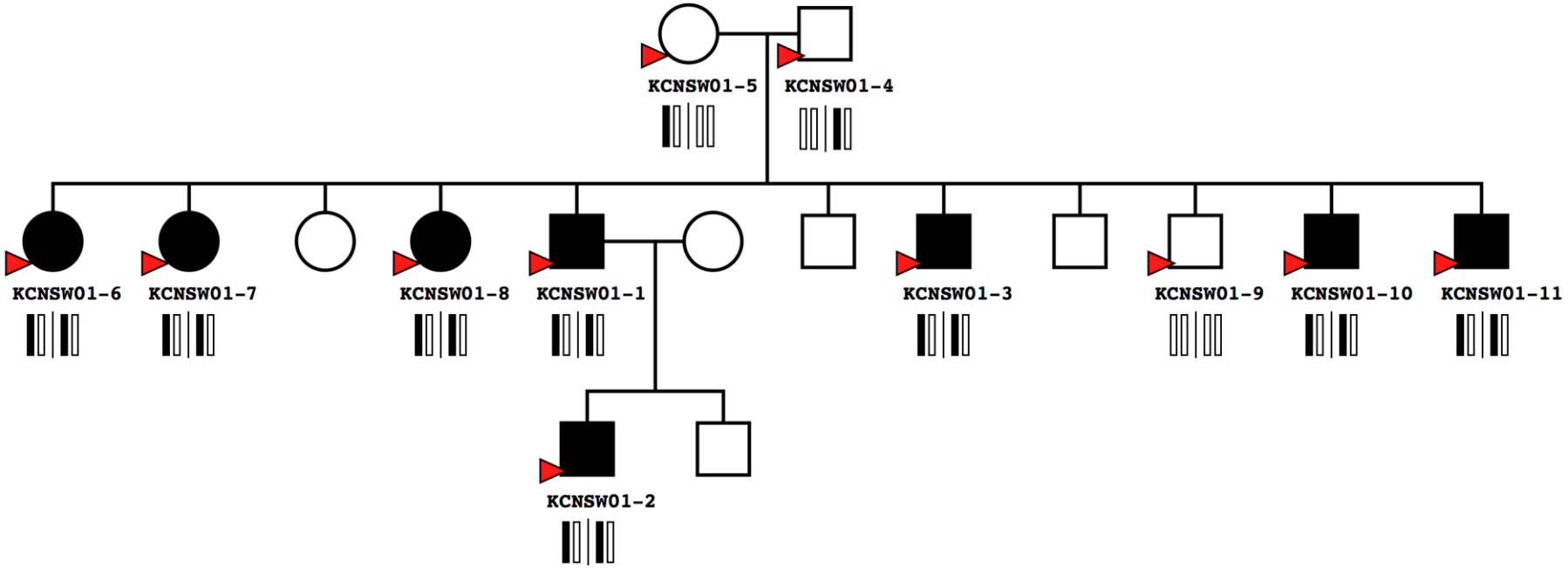


Figure 4.9 – The KCNSW01 pedigree with the addition of each individual's haplotypes at 17q12 and 20p13-12.2.
Females are indicated by circles and males by squares. These symbols are coloured black for individuals with keratoconus and white for unaffected individuals. Red arrows indicate that DNA was collected and WGS data was generated. Haplotypes are shown below the individual ID, with 17q12 on the left of the central bar and 20p13-12.2 on the right. Risk-associated haplotypes are coloured black and all other haplotypes are white. All individuals with keratoconus carry both risk-associated haplotypes.

4.5.3.3 Coverage across the linkage regions

For the chromosome 17 linkage region, a mean depth of 34.1 reads (sd = 3.1) was obtained across the eleven sequenced KCNSW01 family members. Within this region, 649 bases obtained a mean depth below 10 (0.04%). Similarly, the chromosome 20 linkage region obtained a mean depth of 34.6 (sd = 4.0) across the eleven individuals. A total of 43,133 bases within this region had a mean depth below 10 reads (0.42%). This included a 25,521 bp region located at chr20:1561108-1586628.

4.5.3.4 Identifying putatively disease-causing variants

A total of 3,023 variants segregated with the two disease-associated regions identified in KCNSW01, 17q12 and 20p13-12.2. These variants are summarised in Table 4.6.

The autosomal dominant hypothesis

Filtering Strategy 1

A total of 23 protein-coding variants were identified across the two linkage regions (Table 4.7), however, none fulfilled the criteria for putatively disease-causing under Filtering Strategy 1. All but one of the exonic variants were common in at least one population in gnomAD. The single rare variant (rs370777154), located in the zinc finger gene, *ZNF830*, was a synonymous variant and was not predicted to be likely deleterious/pathogenic using either CADD or FATHMM.

Table 4.6 – A summary of number of the segregating variants identified in KCNSW01.

	Variant Type				Variant Location				Highly Prioritised Variants				
Region	**SNPs**	**Ins.**	**Del.**	**Total**	**Exonic (NC)**	**Intronic (NC)**	**UTRs**	**Intergenic**	**Novel**	**Rare**	**CADD ≥ 10 || FATHMM ≥ 0.5**	**FS1**	**FS2**
chr17:32949698-34461869	255	28	23	306	7 (2)	80 (3)	5	214	2	13	21	0	3
chr20:1232032-11479234	2,201	222	294	2,717	42 (24)	1,237 (257)	46	1,392	44	172	216	0	8

SNPs = the number of single nucleotide polymorphisms.
Ins. = the number of insertions.
Del. = the number of deletions.
NC = indicates the total number of non-coding RNA variants included in the variant counts for exonic and intronic variants. Exonic and intronic variants include variants located in protein coding genes as well as non-coding RNA genes.
UTRs = the number of variants located within untranslated regions (either 3' or 5').
Novel = the number of variants without minor allele frequencies in any of the populations in the gnomAD database, Kaviar or IKGP (August 2015 release).
Rare = the number of variants with a minor allele frequency (MAF) less than 1% in all populations of the gnomAD database.
CADD ≥ 10 || FATHMM ≥ 0.5 = the number of variants with a scaled CADD score of at least 10 and/or a FATHMM-MKL/FATHMM-indel score of at least 0.5. This filter is reminiscent of the digenic hypothesis although isn't limited to genes with known interactions.
FS1 = the number of variants that fulfilled the criteria for putatively disease-causing variants under Filtering Strategy 1 (rare protein-coding).
FS2 = the number of variants that fulfilled the criteria for putatively disease-causing variants under Filtering Strategy 2 (not limited to protein-coding).

Table 4.7 – Exonic variants that segregated with the disease-associated haplotypes identified in KCNSW01.

Position	**Ref**	**Alt**	**Gene**	**Variant**	**SNP ID**	**gnomAD max**	**gnomAD NFE**	**CADD**	**FATHMM**
chr17:33288912	**C**	**T**	***ZNF830***	**c.C327T; p.(D109D)**	**rs370777154**	**0.0010**	**0.0001**	**6.61**	**0.01**
chr17:33998802	G	C	*AP2B1*	c.G2061C; p.(V687V)	rs1049379	0.5286	0.5135	5.89	0.90
chr17:34416537	G	A	*CCL3*	c.C180T; p.(P60P)	rs1130371	0.3486	0.2409	11.63	0.02
chr17:34432663	A	G	*CCL4*	c.A237G; p.(E79E)	rs1049807	0.3180	0.2362	0.38	0.01

Position	Ref	Alt	Gene	Variant	SNP ID	gnomAD max	gnomAD NFE	CADD	FATHMM
chr17:34432664	T	A	*CCL4*	c.T238A; p.(S80T)	rs1719152	0.2909	0.2355	$1x10^{-3}$	0.03
chr20:1538266	G	A	*SIRPD*	c.C34T; p.(L12L)	rs11697395	0.3013	0.2268	4.18	0.03
chr20:1546805	G	A	*SIRPB1*	c.C1193T; p.(A398V)	rs62623705	0.2173	0.1462	17.08	0.01
chr20:1552457	G	A	*SIRPB1*	c.C660T; p.(D220D)	rs16995332	0.2185	0.1512	6.91	<0.01
chr20:1600524	T	C	*SIRPB1*	c.A67G; p.(R23G)	rs1535882	0.4522	0.3570	0.09	0.01
chr20:2082767	G	A	*STK35*	c.G240A; p.(Q80Q)	rs6106228	0.8360	0.0358	4.62	0.11
chr20:2291722	A	C	*TGM3*	c.A487C; p.(I163L)	rs6048066	0.3569	0.0013	22.90	0.95
chr20:2638579	T	C	*NOP56*	c.T1424C; p.(M475T)	rs6753	0.5534	0.2762	<0.01	0.02
chr20:2638880	G	A	*NOP56*	c.G1725A; p.(A575A)	rs61752514	0.0220	$6.67x10^{-5}$	6.94	0.06
chr20:2638882	T	C	*NOP56*	c.T1727C; p.(V576A)	rs5856	0.5534	0.2782	<0.01	0.01
chr20:2945759	C	T	*PTPRA*	c.C326T; p.(P109L)	rs1178027	0.5708	0.1038	15.61	0.50
chr20:3452041	C	T	*ATRN*	c.C287T; p.(A96V)	rs75653676	0.0579	0.0003	25.00	0.56
chr20:4768282	G	A	*RASSF2*	c.C810T; p.(Y270Y)	rs6052876	0.1469	0.0002	5.06	0.27
chr20:5903067	T	A	*CHGB*	c.T277A; p.(S93T)	rs6085324	0.3052	0.3775	0.01	0.18
chr20:5903323	G	A	*CHGB*	c.G533A; p.(R178Q)	rs910122	0.3932	0.5835	<0.01	<0.01
chr20:5903848	C	G	*CHGB*	c.C1058G; p.(A353G)	rs236152	0.3949	0.6281	<0.01	0.03
chr20:5903894	A	G	*CHGB*	c.A1104G; p.(E368E)	rs236153	0.3948	0.6286	0.99	0.01
chr20:5904040	G	A	*CHGB*	c.G1250A; p.(R417H)	rs742711	0.3047	0.3776	<0.01	<0.01
chr20:9288522	G	A	*PLCB4*	c.G61A; p.(A21T)	rs6077510	0.6613	0.7020	16.27	0.96

Ref = reference allele.
Alt = alternate allele.
gnomAD max = the maximum alternate allele frequency observed across all populations in the gnomAD database.
gnomAD NFE = the alternate allele frequency in the non-Finnish European population of gnomAD.
CADD = the scaled CADD score.
FATHMM = FATHMM-MKL score for noncoding variants or FATHMM-indel for insertions or deletions.
The rare variant is **bold**.

Filtering Strategy 2

Eleven variants fulfilled the criteria for putatively disease-causing under Filtering Strategy 2, with three SNPs identified on 17q12 and eight variants at 20p13-12.2 (Table 4.8).

In terms of high CADD and FATHMM-MKL scores, the top variant located in the 17q12 region is rs146598068. This appears to be largely driven by the highly conserved nature of the nucleotide. The rs188934690 variant is located in a DNaseI hypersensitive region identified in retinoblastoma tissue, which is cancerous eye tissue derived from the ectodermal lineage. An intronic SNP, rs550832436, is located within a weak enhancer in both epidermal keratinocytes (NHEK cells) and mammary epithelial cells (HMEC cells), however the alternate allele (T) is the reference allele in mouse and rat.

Two variants within the 20p13-12.2 region obtained CADD scores above 20: a novel SNP (chr20:8327708G>A) located in an intron of the gene encoding phospholipase C beta 1 (*PLCB1)* and a previously reported SNP, rs145738299, which is located between a non-coding RNA gene, *PDYN-AS1* and the gene encoding serine/threonine kinase 35 (*STK35*). Both of these variants are highly conserved, however, the alternate alleles observed in this family are the reference alleles in other vertebrate species: the A allele at the novel SNP is observed in aardvark and opossum and the T allele at rs145738299 is present in pacific walrus. Another notable variant located between *MIR8062 – HAO1*, rs140651266, obtained the highest FATHMM score across both regions with a score of 0.99. This variant is located in a DNaseI Hypersensitivity cluster identified in 16 different tissues, including skin fibroblasts and melanocytes, conjunctival fibroblasts (eye tissue derived from the ectoderm), as well as, epithelial and connective tissues.

The most compelling variant identified under Filtering Strategy 2 was a variant located in the 5' untranslated region (UTR) of the spermine oxidase gene, *SMOX*. The *SMOX* variant, c.-224C>T, occurs at the second transcribed nucleotide in the first (entirely untranslated) exon of the gene. This variant is positioned within a DNaseI Hypersensitivity cluster in 120 of the 125 cell-types assessed in ENCODE and overlaps 18 ChIP-seq peaks from the transcription factor ChIP-seq data for 161 factors from ENCODE. This variant is also rare with a maximum allele frequency of 0.38 %, observed in the African population of gnomAD.

Table 4.8 – Putatively disease-causing variants identified under the autosomal dominant hypothesis in KCNSW01.

Position	Ref	Alt	Nearest Gene(s)	SNP ID	gnomAD max	gnomAD NFE	CADD	FATHMM
chr17:33012429-33012429	G	A	*TMEM132E – CCT6B*	rs188934690	0.0081	0.0029	10.84	0.95
chr17:33068005-33068005	G	A	*TMEM132E – CCT6B*	rs146598068	0.0031	0.0021	21.70	0.97
chr17:33351771-33351771	C	T	*RFFL/RAD51L3-RFFL* (intronic)	rs550832436	0.0010	$6.00x10^{-4}$	15.24	0.25
chr20:2007424-2007424	-	G	*PDYN-AS1 – STK35*	novel	0.0033	0.0000	15.05	0.00
chr20:2009593-2009593	C	T	*PDYN-AS1 – STK35*	rs145738299	0.0093	0.0000	22.20	0.97
chr20:2617106-2617115	ACACACACAC	-	*TMC2* (intronic)	novel	0.0000	0.0000	17.23	0.00
chr20:3398935-3398935	T	G	*C20orf194 – ATRN*	rs112827558	0.0023	0.0000	15.35	0.78
chr20:4129427-4129427	**C**	**T**	***SMOX* (c.-224C>T)**	**NA**	**0.0038**	**$2.00x10^{-4}$**	**15.07**	**0.28**
chr20:5945857-5945857	T	C	*MCM8* (intronic)	NA	0.0010	0.0000	12.46	0.97
chr20:7497499-7497499	T	C	*MIR8062 – HAO1*	rs140651266	0.0088	0.0000	16.73	0.99
chr20:8327708-8327708	G	A	*PLCB1* (intronic)	novel	0.0000	0.0000	20.90	0.96

Ref = reference allele.
Alt = alternate allele.
gnomAD max = the maximum alternate allele frequency observed across all populations in the gnomAD database.
gnomAD NFE = the alternate allele frequency in the non-Finnish European population of gnomAD.
CADD = the scaled CADD score.
FATHMM = FATHMM-MKL score for non-coding variants or FATHMM-indel for insertions or deletions.
NA = this variant wasn't attributed an ID in the dbsnp 147 database.
The variant identified in the 5' UTR of *SMOX* variant is in **bold**.

The digenic hypothesis

Ingenuity Pathways Analysis identified protein-protein interactions between the genes that encode chaperonin-containing T-complex polypeptide 1, subunit 6B (*CCT6B)* and *SMOX*,[199] as well as, DNA ligase 3 (*LIG3)* and NSFL1 cofactor (*NSFL1C)*.[200] Additionally, there was evidence that the chemokine gene, *CCL5,* expression is influenced by mitochondrial antiviral signalling protein (*MAVS*),[201-205] signal regulatory protein alpha (*SIRPA*)[206] and prion protein (*PRNP)*.[207] No protein-coding variants located within any of these eight genes segregated with the disease-associated haplotypes and thus segregating, non-protein-coding variants located within the gene or the flanking intergenic regions were further investigated (Table 4.9).

For the genes located within the 17q12 linkage region, seven variants were identified downstream of *CCT6B* (between *TMEM132E* and *CCT6B*), but no variants were identified within the non-coding regions in or around *LIG3* or *CCL5*. As no variants were identified in or around *LIG3* or *CCL5*, bioinformatic investigations focussed on the *SMOX* and *CCT6B* variants, as these genes were the only pair of interacting genes likely to contribute to keratoconus in this family under the present hypothesis. The 5' UTR *SMOX* variant previously identified in Filtering Strategy 2 for the autosomal dominant hypothesis was the only *SMOX* variant identified. As this variant was rare in all populations of gnomAD, it could therefore occur in combination with either a rare or common variant under the digenic hypothesis. Of the seven variants identified downstream of *CCT6B* (*TMEM132E – CCT6B)*, three are the major allele in the non-Finnish European population of gnomAD with alternate frequencies above 50%, two variants rs35933743 and rs4795019 were common with maximum MAFs of 24% and 35% (respectively), and two of the variants were rare (rs188934690 and rs146598068).

All three variants near *CCT6B* with alternate frequencies above 50% – rs1006840, rs1860201, and rs1860200 – are eQTLs for the non-coding RNA gene, *RP11-1094M14.8*, in skin (not sun exposed). The rs35933743 variant is located in a DNaseI hypersensitivity cluster observed in six cells types across all three tissue lineages (mesoderm, ectoderm and endoderm) and the other common variant, rs4795019, is located within a weak enhancer in embryonic stem cells derived from the inner cell mass (H1-hESC). The two rare variants, rs188934690 and rs146598068, were both also identified as putatively causative variants under the autosomal dominant hypothesis. The rs146598068 variant has the highest estimates of deleteriousness/pathogenicity for the *CCT6B* variants, largely owing to the high level of conservation of the nucleotide. This variant is also located within a DNaseI hypersensitivity cluster in astrocytes derived from the hippocampus. As outlined previously, rs188934690 is within a DNaseI hypersensitive cluster observed in a cancerous eye tissue.

Table 4.9 – Putatively disease-causing variants identified in KCNSW01 under the digenic hypothesis.

Position	Ref	Alt	Nearest Gene(s)	SNP ID	gnomAD max	gnomAD NFE	CADD	FATHMM
chr17:32970484	**T**	**G**	***TMEM132E – CCT6B***	**rs35933743**	**0.2404**	**0.1367**	**12.97**	**0.11**
chr17:32998000	**T**	**G**	***TMEM132E – CCT6B***	**rs4795019**	**0.3474**	**0.2271**	**19.62**	**0.97**
chr17:33012429	**G**	**A**	***TMEM132E – CCT6B***	**rs188934690**	**0.0081**	**0.0029**	**10.84**	**0.95**
chr17:33057079	**G**	**A**	***TMEM132E – CCT6B***	**rs1006840**	**0.6216**	**0.5222**	**10.14**	**0.09**
chr17:33057467	**A**	**G**	***TMEM132E – CCT6B***	**rs1860201**	**0.6182**	**0.5186**	**16.54**	**0.15**
chr17:33057477	**C**	**A**	***TMEM132E – CCT6B***	**rs1860200**	**0.6156**	**0.5177**	**15.13**	**0.17**
chr17:33068005	**G**	**A**	***TMEM132E – CCT6B***	**rs146598068**	**0.0031**	**0.0021**	**21.70**	**0.97**
chr20:1422862	T	C	*NSFL1C* (UTR3; c.*1,532A>G)	rs116233763	0.0199	0.0001	11.41	0.20
chr20:1796729	A	G	*LOC100289473 – SIRPA*	rs113396592	0.0733	0.0107	14.49	0.11
chr20:1798513	C	G	*LOC100289473 – SIRPA*	rs8120497	0.0795	0.0105	10.10	0.81
chr20:1802552	A	G	*LOC100289473 – SIRPA*	rs76388170	0.0828	0.0111	16.45	0.18
chr20:1914782	T	C	*SIRPA* (intronic)	rs6045522	0.4372	0.2971	7.918	0.83
chr20:3822099	A	G	*AP5S1 – MAVS*	rs139856798	0.0533	0.0169	10.16	0.30
chr20:3851951	T	C	*MAVS* (UTR3; c.*5,157T>C)	rs6515831	0.4735	0.4586	11.70	0.48
chr20:4129427	**C**	**T**	***SMOX* (UTR5; c.-224C>T)**	**NA**	**0.0038**	**0.0002**	**15.07**	**0.28**
chr20:4288521	T	C	*ADRA1D – PRNP*	rs8117034	0.1806	0.0302	10.21	0.35
chr20:4296669	A	-	*ADRA1D – PRNP*	rs142211335	0.1814	0.0304	19.41	0.01
chr20:4296675	A	G	*ADRA1D – PRNP*	rs116181036	0.1813	0.0302	17.02	0.27
chr20:4306646	G	A	*ADRA1D – PRNP*	rs4815687	0.5951	0.0884	10.92	0.85
chr20:4403389	G	T	*ADRA1D – PRNP*	rs75400769	0.0176	0.0001	5.796	0.68
chr20:4426668	C	G	*ADRA1D – PRNP*	rs79099834	0.0226	0.0001	11.98	0.26
chr20:4436181	C	T	*ADRA1D – PRNP*	rs115014683	0.0220	0.0001	7.358	0.51
chr20:4451485	A	G	*ADRA1D – PRNP*	rs16989700	0.0590	6.66×10^{-5}	12.23	0.21

chr20:4485701	A	G	*ADRA1D – PRNP*	rs60514595	0.2031	0.0014	13.31	0.09
chr20:4493690	C	T	*ADRA1D – PRNP*	rs73893468	0.0504	0.0003	12.90	0.69
chr20:4502865	A	G	*ADRA1D – PRNP*	rs374847600	0.3372	0.0036	11.11	0.35
chr20:4509265	A	C	*ADRA1D – PRNP*	rs11906771	0.1714	0.0005	10.74	0.57
chr20:4509511	T	-	*ADRA1D – PRNP*	rs141573094	0.0614	0.0003	11.52	0.00
chr20:4523958	T	C	*ADRA1D – PRNP*	rs73893488	0.0628	0.0003	11.33	0.40
chr20:4567204	A	G	*ADRA1D – PRNP*	rs11906818	0.0720	0.0005	7.333	0.87
chr20:4634737	A	T	*ADRA1D – PRNP*	rs113394511	0.1671	0.0504	10.36	0.17
chr20:4644408	T	C	*ADRA1D – PRNP*	rs74799977	0.2951	0.1997	10.95	0.39
chr20:4646394	T	A	*ADRA1D – PRNP*	rs75731019	0.0522	0.0001	10.53	0.11
chr20:4648760	TT	-	*ADRA1D – PRNP*	rs796775108	0.0538	0.0023	11.12	0.00
chr20:4652694	A	G	*ADRA1D – PRNP*	rs73896119	0.1190	0.0175	8.63	0.69
chr20:4652735	A	C	*ADRA1D – PRNP*	rs16989990	0.1193	0.0176	15.82	0.11

Ref = reference allele.
Alt = alternate allele.
Position = the position of SNPs or the start position for indels.
gnomAD max = the maximum alternate allele frequency observed across all populations in the gnomAD database.
gnomAD NFE = the alternate allele frequency in the non-Finnish European population of gnomAD.
CADD = the scaled CADD score.
FATHMM = FATHMM-MKL score for non-coding variants and FATHMM-indel scores for indels.
NA = this variant wasn't attributed an ID in the dbsnp 147 database.
Variants in *CCT6B* and *SMOX* are **bold**.

4.6 DISCUSSION

This study identified two novel regions linked to keratoconus and replicated a third: a single homozygous region on 16p12.1 was identified in the two affected brothers in a family with consanguinity (KSA197), and two regions (17q12 and 20p13-12.2) showed equal evidence for linkage to keratoconus in a family displaying apparent autosomal dominant disease with reduced penetrance (KCNSW01). The inheritance pattern of the two disease-associated haplotypes throughout the KCNSW01 pedigree was strongly indicative of digenic inheritance, with all affected individuals carrying both disease-linked regions. This suggests that each disease-associated haplotype may harbour a single variant that only cause keratoconus when inherited together, with neither sufficient to cause disease in isolation.

KCNSW01 is the second family demonstrating strong evidence of digenic inheritance of keratoconus. Another family in our Australian keratoconus cohort previously showed likely digenic inheritance of keratoconus with parametric linkage analysis identifying two suggestive linkage regions with equal LOD scores: 1p36.23-36.21 and 8q13.1-q21.11.[71] These regions reached significance when analysed together under a digenic model (LOD = 3.4) and all family members with keratoconus carried both disease-associated haplotypes.[71] Similarly, all affected family members in KCNSW01 in the present study carried both disease-associated haplotypes at 17q12 and 20p13-12.2, while each parent who carried only one was reported as unaffected. The 20p13-12.2 linkage region identified in KCNSW01 is completely encompassed by a previously published linkage region identified in another family in which two suggestive regions were identified.[66] The previous study identified suggestive linkage regions at 2q13-14.3 and 20p13-12.2 (LOD = 2.4) in a large Ecuadorian family, however unlike KCNSW01, the inheritance pattern of the disease-associated haplotypes did not clearly indicate digenic inheritance.[66] In the Ecuadorian family the inheritance pattern of the disease-associated haplotype on 2q13-14.3 suggests autosomal dominant inheritance with reduced penetrance, with all affected individuals and two unaffected individuals carrying the risk-associated haplotype, while the segregation of the disease-associated haplotype at 20p13-12.2 is less clear with only eight of the nine cases sharing the same haplotype, along with three individuals who are unaffected or have an unknown phenotype. In both the Ecuadorian family and KCNSW01, it is possible that the 20p13-12.2 region showed suggestive association with keratoconus by chance, however the fact that the same region was identified in two unrelated families – each with a second, equally associated region – is noteworthy. This adds evidence for the involvement of variation within this region in keratoconus susceptibility. Potentially, this region harbours a variant with an intermediate effect size, and a second variant is required to trigger keratoconus development.

Prioritising variants in keratoconus is challenging, even in families with strong Mendelian inheritance patterns such as KSA197 and KCNSW01, as little is known about the types of genes or variants

involved in the disease. This is made even more challenging in non-coding portions of the genome due to our limited understanding of these regions in normal biology. To aid the prioritisation of variants within the discovered regions of interest, two *in silico* prediction tools were used in combination: CADD[134] and FATHMM[135]. Both of these tools aggregate multiple annotations (including nucleotide conservation and known regulatory regions) into a single metric predicting deleteriousness or pathogenicity and are commonly used to help differentiate between variants that are likely to be functional and those that are likely to be benign. Scaled CADD scores above 10 correspond to the top 10% most deleterious substitutions possible in the human genome and scores above 20 correspond to the top 1% and so on.[134] The scores from the FATHMM algorithms range from 0 to 1 with values of 0.5 or above predicted to be deleterious and the higher the score, the greater the confidence that the variant is functional.[136] CADD is considered the best tool for detecting pathogenic variants in protein coding regions of the genome, while the FATHMM-MKL algorithm outperforms CADD and other commonly used *in silico* tools for the pathogenicity prediction of variants located in the non-coding regions.[208] Based on this, both tools were used in combination to limit the exclusion of true disease-causing variation. Unlike the previous generation of pathogenicity prediction tools such as SIFT[132] and PolyPhen2,[133] CADD and FATHMM score both coding and non-coding variation, as well as, indels (FATHMM uses an indel specific algorithm: 'FATHMM-indel'), allowing for consistency across WGS data.

Two separate hypotheses were used to interrogate putatively disease-causing variation within the two regions of interest in KCNSW01. The autosomal dominant hypothesis was used to identify rare variants with high predictions of deleteriousness or pathogenicity under the assumption that only one variant was required to cause keratoconus in KCNSW01. Two filtering strategies were used to identify putatively disease-causing variants under this hypothesis, with Filtering Strategy 1 limiting the analysis to protein-coding variants and Filtering Strategy 2 including non-coding variation. Both of these strategies required the putatively disease-causing variants to obtain high estimates of pathogenicity or deleteriousness a scaled CADD score of at least 15 and/or a FATHMM-MKL/FATHMM-indel score of 0.8. This CADD threshold was suggest by the authors for identifying potentially pathogenic variants as it is the median value for all possible splice site changes and nonsynonymous variants.[134] Commensurate with the autosomal dominant hypothesis, a threshold of 0.8 was applied to the FATHMM algorithms to identify likely deleterious variants. No putatively disease-causing variants were identified under Filtering Strategy 1, suggesting that rare protein-coding variants are unlikely to be the cause of keratoconus in this family. In contrast, Filtering Strategy 2 identified three within the 17q12 linkage region and eight in the 20p13-12.2 region. Conversely, the digenic hypothesis assumed that both linkage regions harbour a variant, that only cause keratoconus in combination and thus focused on variants located within or near pairs of genes (one from each of the regions of interest) with documented interactions. Under this hypothesis, putatively disease-causing variants were identified

using lower thresholds for deleteriousness and pathogenicity – a scaled CADD score of 10 or above and/or FATHMM score of at least 0.5 – to account for the fact that these variants are not sufficient to cause disease in isolation. Moreover, no minor allele frequency threshold was applied pertaining to the fact that the two variants only need to be rare in combination. Putatively disease-causing variants were identified at both gene partners in only one pair of interacting genes: *SMOX* and *CCT6B*. A single putatively-disease causing variant was identified at *SMOX* and seven were identified within the non-coding region downstream of *CCT6B.*

The *SMOX* variant, c.-224C>T, was identified within the 20p13-12.2 linkage region of KCNSW01. This variant was classified as putatively disease-causing under both hypotheses used to prioritise variation in this family. *SMOX* encodes the enzyme spermine oxidase that catalyses the oxidation of spermine to spermidine.[209] This reaction also produces the reactive oxygen species, H_2O_2, and the reactive aldehyde, aldehyde 3-aminopropanal.[209] Spermine and spermidine are both polyamines, are found ubiquitously in all organisms and play an important role in a number of cellar processes including protein translation, response to oxidative stress and ultraviolet radiation and apoptosis.[210] Spermine oxidase is inducible by inflammation, as studied in bronchial epithelial cells[211] and gastric epithelia,[212, 213] and has a role in pathologies such as gastritis and epithelial cancers through the induction of DNA damage and apoptosis. While keratoconus has long been reported as a non-inflammatory disease, studies in the last two decades have demonstrated a likely role of chronic inflammation in keratoconus, with significant increases in a number of inflammatory mediators such as interleukin-6,[214-217] and significantly lower anti-inflammatory markers such as IL-10[214] and CCL5[217] in tears from keratoconus patients compared to controls. Keratoconus patients have also been shown to have increased systemic oxidative stress levels compared to age matched healthy controls.[218] There is also evidence of the association of keratoconus with increased apoptosis, with histopathological studies of keratoconic corneas demonstrating marked increases in DNA fragmentation[219, 220] and single stranded DNA,[219] both markers for apoptosis, compared to healthy control corneas. Taken together, *SMOX* is a strong candidate gene for keratoconus and both the variant identified in KCNSW01, and the gene, warrant functional investigation.

As the *SMOX* variant (c.-224C>T) alters the second nucleotide in the 5' UTR, we hypothesise that this variant will result in an altered transcription start site and ultimately abnormal expression levels of the encoded protein and that this contributes to keratoconus susceptibility in KCNSW01. Examples of causal variants located within the 5' UTR of key proteins have been identified for a number of monogenic disorders such as Saethre-Chotzen syndrome[221] (premature fusion of the cranial sutures) and androgen insensitivity syndrome,[222] as well as a severe immune deficiency resulting in chronic mycobacterial infections.[223] There has also been an example of two 5' UTR variants that cause isolated congenital asplenia with incomplete penetrance, [224] a condition in which children are born without a spleen in the absence of any other developmental defects caused by haploinsufficiency of a ribosomal

protein (RPSA). This study demonstrated that both 5' UTR variants impaired splicing at the exon 1/intron 1 junction, resulting in the retention of 67 or 70 nucleotides and introducing a second potential translation start site. The resulting mRNA molecule was stable, but depending on the translation start site utilised, potentially resulted in a mutant protein. From this work, the authors hypothesised that the ratio of mutant and wild-type protein produced determined whether or not the phenotype was observed and therefore described the variants as 'hypomorphic' as they caused smaller decreases in the functional protein than other well characterised heterozygous variants. As the *SMOX* variant is present in the apparently unaffected patriarch of KCNSW01, and variable keratoconus severity is reported across the affected family members, a similar mechanism could explain the heterogeneity of disease in this family. Furthermore, due to the likely digenic inheritance in KCNSW01, potential interactions between the *SMOX* variant and the variants downstream of *CCT6B*, or other variants within the 17q12 linkage region, may be modifying the phenotype within the family. Functional studies are required to confirm the role of this variant in disease, as well as, determine the mechanism of disease.

A single putatively disease-causing variant (chr16:27115706C>A) was identified within the homozygosity region identified on chromosome 16 in KSA197 under Filtering Strategy 2. No rare protein-coding variants were identified within the homozygosity region and therefore Filtering Strategy 1 did not identify any protein-coding putatively disease-causing variants. Due to the inheritance pattern and consanguinity in KSA197, it was hypothesised that the causative variant was homozygous in the affected individuals and that the variant in the heterozygous state was not sufficient to cause disease. For this reason, the same low CADD and FATHMM thresholds as those used under the digenic hypothesis for KCNSW01 were applied during the prioritisation of variants in KSA197. Putatively disease-causing variants in KSA197 were also required to be rare (MAF < 1%). The putatively disease-causing variant identified (chr16:27115706C>A) was a novel SNP located between the uncharacterised, single exon gene, *C16orf82*, and a histone lysine demethylase gene, *KDM8*. Pathogenicity estimates are highly suggestive of the functionality of this variant. This variant is also located within a DNaseI hypersensitivity region identified in osteoblasts, suggesting that this region may have a regulatory role in certain tissues and stages of development. Functional studies in relevant ocular cell types are required determine the potential role of this variant in keratoconus.

The human chromosome 16 contains more repetitive sequence than the other human chromosomes, largely due to low-abundance repetitive sequences that uniquely occur on this chromosome.[225] Clusters of these sequences were identified during the construction of a physical map of chromosome 16, including 16p12 which overlaps the homozygosity region identified in KSA197.[225] Repetitive DNA sequences are difficult to call from short-read sequencing as they create ambiguities in the alignment of reads and can lead to incorrect variant calls, including spurious indels.[226] This explains why many indels located at repetitive DNA sequences within the homozygosity region were called in the heterozygous state in one or both of the affected brothers, who should have been homozygous. In the present study,

these variants were excluded from analysis as they didn't segregate with disease, however it is possible that these repetitive sequences are relevant to keratoconus-susceptibility in KSA197. Repetitive DNA sequences are known to be involved in repeat-expansion disorders such as Huntington's chorea[227] and fragile X syndrome,[228, 229] but have also been shown to increase susceptibility to disease such as insulin-dependent diabetes mellitus.[230, 231] Nearly 30 disorders have been linked to repetitive sequences within the coding portions of genes or nearby non-coding regions,[232] and therefore, the repeat regions within the homozygosity region identified on chromosome 16 in KSA197 should be further investigated. It would be worthwhile exploring short tandem repeat (STR) sequences – which consist of 1-6 bp sequences repeated one after another, typically 5-50 times – within the homozygosity region. A recently developed software, STRetch,[233] has been designed to detect STR sequences and estimate their size, allowing for the identification of pathogenic repeat expansions using short-read WGS data. This method could be applied to our current WGS data. It is also possible to obtain long-read sequencing data which typically generates reads greater than 5 kbp, although this would incur a considerable cost. While long-read sequencing has high error rates,[234] these data would be used as a scaffold for our short-read WGS data and together this would improve our ability to accurately call variants, allow us to investigate complex structural variation and thoroughly explore the repeat sequences within the homozygosity region.

All putatively disease-causing variants identified in KCNSW01 and KSA197 in the present study were non-protein-coding variants. In fact, only one protein-coding variant segregated with disease within the homozygosity region identified in KSA197, but this variant was too common in the general population to be the causative variant. This finding is consistent with previous linkage studies which have struggled to identify strong candidate variants within coding regions of genes located within keratoconus linkage regions. As a complex disease, it is quite likely that more complex genetic mechanisms are involved in keratoconus, such as non-coding variants in key regulatory regions. This hypothesis is supported by the identification of the *SMOX* 5' UTR variant in KCNSW01 in the present study as well as other candidate genes proposed in previous linkage studies for keratoconus. The gene dedicator of cytokinesis 9 (*DOCK9*) was proposed as a candidate gene for keratoconus following the identification of a segregating splicing variant in an Ecuadorian family that results in exon skipping and the incorporation of a premature stop codon in the gene transcript prior to the functional domains of the protein.[29, 120] Intronic variants in *IL1RN* and *SLC4A11* are hypothesised to alter the expression of these proteins were proposed to play a role in keratoconus development following linkage analysis in another Ecuadorian family.[66] Finally, the recurrent variant associated with a combined keratoconus and congenital cataract phenotype is located in the non-protein-coding RNA gene, *mir184.*[65, 69, 77, 143] This gene is highly expressed in the cornea and lens and is hypothesised to cause the phenotype by altering the expression of key proteins by binding to their mRNA molecules with altered affinity compared to the wildtype mir184.[65] While these variants need to be further assessed to confirm their involvement in keratoconus

and determine the exact mechanism of disease, together they suggest that non-protein-coding variation is likely to be important in keratoconus development. Perhaps one of the key reasons that specific keratoconus variants have been so elusive is precisely because the field has largely focused on protein-coding variation in families with multiple cases of disease.

A key limitation of the study was access to clinical information for KCNSW01 as the vast majority of the family are living in Jordan and could not be assessed by our ophthalmologists. Despite this the almost complete ascertainment of a family as large as KCNSW01, particularly when family members are living internationally, is incredibly difficult and therefore represents one of the strengths of the study. Furthermore, the severity of the disease reported in the affected individuals gave us confidence for the phenotype assignment in these individuals. It is however possible that one or both of the parents in the first generation, and the single individual in the second, may have mild or subclinical keratoconus. As the inheritance pattern in the second and third generations was consistent with autosomal dominant disease, it was a concern that both the matriarch and the patriarch were reported unaffected, therefore these individuals were coded with unknown phenotypes and linkage analysis was conducted using a reduced penetrance model. If keratoconus is inherited as a digenic trait as the linkage results and haplotype analysis would suggest, it is however, quite likely that these individuals are unaffected. Despite these limitations, two regions showed segregation with disease in this family, allowing for prioritisation of putatively disease-causing variants. As with almost any family study, the examination and recruitment of additional unaffected family members from both families may refine the regions of interest and aid the prioritisation of candidate variants for further investigation.

The strategic use of WGS data to both conduct linkage in two families and interrogate variants harboured within the identified regions represents an important strength of this study. This method eliminated the need for additional SNP array genotyping and facilitated the interrogation of variation within the linked regions without *a priori* hypothesis. Like any sequencing method, variants located within regions with low coverage may have been overlooked, however, WGS consistently outperforms WES for the capture of the protein-coding portion of the genome as it is not biased by a probe-based capture.[235, 236] The WGS data also allowed for the investigation of non-coding variation, making this study the most comprehensive family study in keratoconus to date, and these data are likely to become more valuable as our knowledge of non-coding variation and specific variants involved in keratoconus development and pathogenesis improve over time. Furthermore, these data present a unique opportunity to explore novel genetic mechanisms in keratoconus susceptibility by investigating structural variation and microsatellites in these families.

4.7 CONCLUSION

This study identified two novel linkage regions for keratoconus (16p12.1 and 17q12), replicated a third (20p13-12.2) across two families, with one of the families (KCNSW01) demonstrating likely digenic inheritance of keratoconus. Despite the original hypothesis, no rare protein-coding variants were classified as putatively-disease causing in either family. When considering non-protein-coding variants that segregated with the disease-associated regions, one putatively disease-causing variant was identified in KSA197 and 44 in KCNSW01 (combining the results of both hypotheses), including a compelling variant located within the 5' UTR of *SMOX*. These variants warrant further investigation in additional patients with keratoconus, as well as functional analyses in relevant tissues or cell lines. Based on the lack of rare, coding putatively disease-causing variants identified in the present study, we hypothesis that non-coding variants are likely to contribute to keratoconus development in these families and potentially play an important role in keratoconus susceptibility and pathogenesis more broadly.

CHAPTER 5: MAPPING PUTATIVELY FUNCTIONAL RISK ALLELES AT KERATOCONUS-ASSOCIATED LOCI

Work outlined in Aim 1 of this chapter was included in the following publication:

Iglesias A*, Mishra A*, Vitart V*, Bykhovskaya Y, Hohn R, Springelkamp H, Cuellar-Partida G, Gharahkhani P, Cooke Bailey J, Willoughby C, Li X, Yazar S, Nag A, Khawaja A, Polasek O, Siscovick D, Mitchell P, Tham YC, Haines J, Kearns L, Hayward C, Shi Y, van Leeuwen E, Taylor K, Bonnemaijer P, Rotter J, Martin N, Zeller T, Mills R, Souzeau E, Staffieri S, Jonas J, Schmidtmann I, Boutin T, **Lucas SEM**, Kang J, Wong T, Beutel M, Wilson J, Vithana E, Foster P, Hysi P, Hewitt A, Khor CC, Pasquale L, Montgomery G, Klaver C, Aung T, Pfeiffer N, Mackey D, Hammond C, Cheng C, Craig JE, Rabinowitz Y, Wiggs J, Burdon KP, Duijn C, MacGregor S. **Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases.** *Nature Communications* 2018;9(1):1864.

* indicates equal contribution.

5.1 INTRODUCTION

In genetics, association studies use cohorts of unrelated individuals to identify variants that contribute to a heritable trait or disease. In recent decades, high throughput DNA genotyping technology has allowed for genome-wide association studies (GWAS), which assess thousands of single nucleotide polymorphisms (SNPs) across the genome for association with a disease without *a priori* hypotheses. To account for multiple testing in these studies, a genome-wide significance threshold is applied (5×10^{-8}). In contrast to family-based studies, association studies are well powered to identify common variation, and as such, the variants identified often have small to moderate effect sizes. In the context of complex disease, this type of analysis identifies variation that confers slight increases or decreases in disease susceptibility. It's important to note that due to linkage disequilibrium (LD) – the co-inheritance of alleles at different loci within a population – an associated variant may directly contribute to the risk of developing the disease or may just be in LD with the true functional variant (assuming that the association is not a false-positive). Therefore, while associations are often referred to by the associated SNP ID, or a nearby gene, in reality the association implicates a specific haplotype. The challenge then remains to identify the functional variant and target gene to gain an insight into the underlying biology and pathogenesis of disease.

Many small case-control studies have assessed variants in and near genes hypothesised to play a role in the keratoconus disease process for association in keratoconus cohorts, including *VSX1*,[89] *CAST*,[113] and *SOD1*,[103] however with such limited understanding of the types of variants and genes involved in

keratoconus, these studies have not yet lead to substantial insights into the pathogenesis of disease. There have also been two published GWAS for keratoconus. The first GWAS for keratoconus was performed using our cohort of Australian keratoconus patients and identified a suggestive association at rs3735520 ($p = 9.9 \times 10^{-7}$) upstream of *HGF*.[84] This association has not been replicated and therefore the involvement of this locus in keratoconus susceptibility remains unclear. The second GWAS identified a suggestive association ($p = 1.6 \times 10^{-7}$) at rs4954218 upstream of *RAB3GAP1*,[85] which reached genome-wide significance following replication and meta-analysis in our cases ($p = 5.0 \times 10^{-8}$).[86]

Three additional loci have since reached genome-wide significance with keratoconus[87] – *FOXO1*, *FNDC3B*, and *MPDZ-NFIB* – however, these associations were identified using an endophenotype approach, rather than a direct GWAS for keratoconus. The concept of an endophenotype has evolved substantially since the term was first used in 1966,[237] but currently refers to a quantitative trait that is associated with the disease in the population, is heritable, is measurable in both healthy and affected individuals, co-segregates within families (like the disease) and found in family members unaffected by the disease at a higher rate than the general population.[238-241] By definition endophenotypes are quantitative, resulting in better statistical power and higher precision when making inferences compared to discrete data (such as affected/unaffected). This allows for smaller sample sizes and more efficient use of resources. Endophenotypes are also hypothesised to be less complex and closer to the underlying genetics than the disorder or disease of interest, as it reflects just one of many pathophysiological pathways that contribute to disease susceptibility.[238] This allows complex diseases to be dissected, directing investigators toward the relevant biological pathways and ultimately aiding the identification of functional variants and target genes. In keratoconus, central corneal thickness (CCT) has been used as an endophenotype to aid the identification of keratoconus-associated loci. CCT is a quantitative trait with a normal distribution in the general population (mean = 536 µm; standard deviation = 31 µm) and is measurable in both healthy and keratoconic corneas.[31] Though the genetic correlation between CCT and keratoconus has not been calculated, keratoconus is associated with extremely low CCT with a mean of 434 µm reported in keratoconic eyes,[31] which is well outside the variance observed in the general population. CCT is one of the most heritable human traits with heritability estimates of up to 95%.[242-245] Unaffected family members of individuals with keratoconus have been shown to have thinner corneal measurements when compared to population controls,[246, 247] with some evidence of an autosomal dominant pattern of inheritance.[247] Taken together, CCT is a strong endophenotype for keratoconus.

The first study that used CCT as an endophenotype for keratoconus performed a GWAS for CCT and subsequently typed genome-wide significant findings in a cohort of keratoconus patients that included our Australian cases and a cohort from the USA.[87] The investigators hypothesised that the CCT-decreasing alleles at CCT-associated loci would also confer an increased risk of keratoconus and that

CCT-increasing alleles may be protective for keratoconus. The study identified significant associations at rs2721051 in an intron of *FOXO1* ($p = 2.7 \times 10^{-10}$) and the intronic SNP rs4894535 in *FNDC3B* ($p = 4.9 \times 10^{-9}$).[87] A suggestive association at rs1324183 between *MPDZ* and *NFIB* ($p = 5.2 \times 10^{-6}$)[87] was also identified in this study and reached genome-wide significance following replication and meta-analysis in an independent cohort ($p = 5.0 \times 10^{-8}$).[88] Notably, suggestive associations at rs1536482 ($p = 2.6 \times 10^{-7}$) between *RXRA-COL5A1* and rs7044529 in an intron of *COL5A1* were also identified.[87] Another study employed a similar study design and identified another suggestive association with keratoconus at rs121908120, located in an exon of *WNT10A* ($p = 5.4 \times 10^{-5}$).[248] Again, our keratoconus cohort was used in this study and replication in an independent cohort is required to determine if this locus is important in keratoconus susceptibility. These studies highlight the success of assessing CCT-associated loci in a keratoconus cohort to aid the identification of novel keratoconus-associated loci.

Finding a significant association between a variant and a disease is only the first step toward a better understanding of the underlying biology and mechanism of disease. More often than not, the associated variants are located in intronic and intergenic regions and substantial investigations are required to elucidate the functional variant and target gene(s). Fine-mapping is a method used to establish the extent of an associated region, as well as identify the 'top SNP', the variant that is most strongly associated with the trait of interest (ie. obtains the smallest p-value). This may involve hard-typing nearby SNPs or, more commonly in the era of GWAS, imputation. Imputation uses known genotypes, commonly from a SNP array, to infer an individual's genotype at additional un-genotyped SNPs, based on known haplotypes such as those from the HapMap Project[191] or 1000 Genomes Project.[131] Following fine-mapping, variants carried on the risk-associated haplotype can subsequently be assessed using *in silico* tools to identify putatively functional variants, which ultimately require functional investigation to confirm the role of the variation in disease susceptibility. To date, the functional variants at keratoconus-associated loci have not yet been investigated.

5.2 HYPOTHESIS AND AIMS

The hypothesis underpinning this study was that variants associated with keratoconus indicate haplotypes that harbour functional variants that directly contribute to disease susceptibility. This led to the development of the following aims:

1. To identify novel keratoconus-associated loci by assessing central corneal thickness-associated loci in keratoconus patients and unaffected controls;
2. To fine-map keratoconus-associated loci in a cohort of unrelated keratoconus cases and controls to investigate the extent of the association, identify the top SNP, and select genomic regions for re-sequencing. This aim will assess novel keratoconus-associated

loci identified in Aim 1, as well as, published keratoconus-associated loci that have reached genome-wide significance.

3. To identify putatively functional variants underlying keratoconus-associated loci by re-sequencing keratoconus patients carrying the risk-associated alleles. This aim will focus on fine-mapped regions with strong association peaks identified in Aim 2.

5.3 OVERALL STUDY DESIGN

The overall objective of this chapter was to investigate the role of common genetic variation in keratoconus susceptibility in a large, unrelated case-control cohort. This study includes three specific aims as outlined in Figure 5.1.

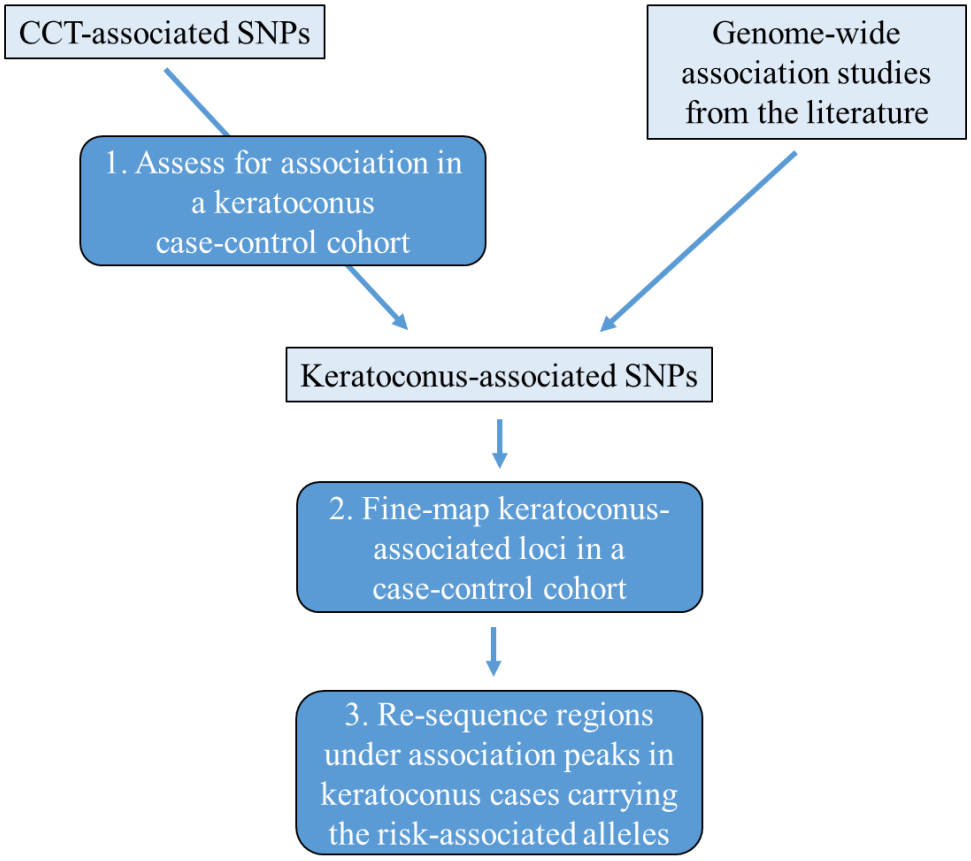


Figure 5.1 – A flow diagram of the overall study design.
Dark blue textboxes indicate the three aims, numbered 1 to 3. CCT = central corneal thickness; SNPs = single nucleotide polymorphisms.

5.4 AIM 1: ASSESSING CENTRAL CORNEAL THICKNESS-ASSOCIATED LOCI IN A KERATOCONUS COHORT

5.4.1 Methods

5.4.1.1 SNP selection

Our team and collaborators conducted a cross-ancestry GWAS for CCT using 20 cohorts, including 17,803 individuals of European descent and 8,107 individuals of Asian ancestry (total n = 25,910). This work identified 54 independent association signals for CCT at 44 loci, including 21 novel loci.[249] Based on this work, 72 CCT-associated SNPs that were novel findings or had not previously been assessed in keratoconus were selected for evaluation in our case-cohort. The results from an LD-pruned list of these SNPs (n = 36) were included in the published study following meta-analysis with a cohort from the United States of America (USA).[249]

5.4.1.1 Study participants

The case cohort consisted of 536 unrelated keratoconus patients as described in Section 2.1.1. The control cohort included 2,574 unaffected individuals from the Blue Mountains Eye Study (BMES). This cohort is outlined in Section 2.1.4.

5.4.1.2 Genotyping

The 72 CCT-associated SNPs were genotyped in cases across three custom plexes by the Australian Genome Research Facility (QLD, Australia) using the MassARRAY platform (Agena Biosciences Inc., San Diego, California, USA).

The control cohort were previously typed on the HumanHap 610 array (Illumina) and custom CoreExome array (Illumina). Quality control and genome-wide imputation was conducted by Dr Puya Gharahkhani (Queensland Institute of Medical Research Berghoefer, QLD, Australia). Using PLINK[190] (version 1.9) SNPs were excluded if they had a call rate less than 97%, a minor allele frequency (MAF) < 0.01 and a Hardy-Weinberg equilibrium $p < 0.0001$. Individuals with more than 3% missing genotypes were excluded. Related individuals were identified by calculating identity-by-descent (IBD) in PLINK using autosomal markers, and one individual from each pair with IBD > 0.2 was removed from analysis. Principal components analysis (PCA) was conducted in PLINK for all individuals and reference samples of known European ancestry from the 1000 Genomes Project: Utah residents with Northern and Western European ancestry (CEU), British living in England and Scotland (GBR) and Finnish living in Finland (FIN) populations.[250] Individuals with PC1 or PC2 values > 6 standard deviations from the mean of the European 1000 Genomes Project reference groups were considered ancestry outliners and excluded from further analysis. Following this quality control protocol, 2,574 individuals and 537,548 SNPs were used as the basis of imputation.

Genotypes were phased using ShapeIT[251] (v2.r790) and imputation was performed using Minimac3[252] (version 2.0.1) through the Michigan Imputation Server with the Haplotype Reference Consortium[253] (release 1.1) as the reference panel. SNPs with imputation quality (r^2) > 0.3 and MAF > 0.01 were available for analysis. The genotype data for the controls at the 72 CCT-associated SNPs were extracted from this dataset.

5.4.1.3 Statistical analysis

Genotype data for the case and control cohorts were merged in PLINK. Individuals with more than 10% missing genotype data were removed from the analysis. SNPs with a missing genotype rate greater than 10% or that deviated significantly ($p < 0.001$) from Hardy-Weinberg Equilibrium were excluded. Using PLINK, chi squared tests with odds ratios and 95% confidence intervals were calculated. An example of the PLINK command is outlined in Appendix 15. To correct for multiple testing Bonferroni correction, a significance threshold of 6.94×10^{-4} was applied (0.05/72).

5.4.2 Results

A total of 72 CCT-associated SNPs were assessed for association with keratoconus using a cohort consisting of 536 keratoconus patients and 2,574 controls. The cohort demographics are outlined in Table 5.1.

Table 5.1 – Cohort demographics

Group	n	Mean age (range)	% Female	Disease status
Cases	536	44.6 (14 – 85)	44.6	affected
Controls	2,574	70.0 (49 – 96)	43.4	unaffected

n = the number of individuals in each cohort.
Mean age = the mean average (and the range) is reported in years.

Hard-typed genotype data were available for all 72 CCT-associated SNPs in the case cohort and 33 of the SNPs in the controls. Imputed genotype data were available for the remaining 39 SNPs in the control cohort. Twenty-five SNPs were nominally associated with keratoconus ($p < 0.05$) and five remained significant following adjustment for multiple testing (Table 5.2). Of the significant SNPs, rs2268578 was located at a novel locus in an intronic region of the lumican gene (*LUM*) and two SNPs between *RXRA* and *COL5A1* (rs1536482 and rs3132303) were significantly associated with keratoconus. This study also replicated two genomic regions that had previously reached genome-wide association with keratoconus: *FOXO1* with the significant association at the intronic SNP rs2755238 and *MPDZ-NFIB* with an association at rs66720556, located between the two genes. Based on this analysis, the novel loci

LUM and *RXRA-COL5A1*, along with the known loci *FOXO1* and *MPDZ,* were further fine-mapped in Aim 2 (Section 5.5).

Table 5.2 – CCT-associated SNPs assessed for association in our cohort of keratoconus patients and unaffected controls.

SNP	Position	Locus	Ref	Alt	Case Freq.	Control Freq.	X^2	P	OR [95% CI]	SE
rs96067*	chr1:36571920	*COL8A2*	A	G	0.22	0.20	2.56	0.11	1.14 [0.97-1.34]	0.08
rs4846476	chr1:218526228	*TGFB2*	G	C	0.23	0.23	0.35	0.55	1.05 [0.90-1.23]	0.08
rs115781177	chr2:33348494	*LTBP1*	A	G	0.06	0.07	3.11	0.08	0.78 [0.59-1.03]	0.14
rs4608502	chr2:228134155	*COL4A3*	C	T	0.38	0.33	8.24	4.11×10^{-3}	1.22 [1.07-1.40]	0.07
rs12469734*	chr2:235469549	*ARL4C*	A	G	0.39	0.36	3.15	0.08	1.13 [0.99-1.30]	0.07
rs9880211*	chr3:136107549	*STAG1*	G	A	0.28	0.24	7.60	5.83×10^{-3}	1.23 [1.06-1.43]	0.08
rs28641809	chr3:136290489	*STAG1*	G	A	0.28	0.23	8.17	4.26×10^{-3}	1.24 [1.07-1.44]	0.08
rs13092225*	chr3:156355404	*TIPARP*	A	G	0.36	0.34	1.62	0.20	1.09 [0.95-1.26]	0.07
rs6807894*	chr3:156372177	*TIPARP*	T	G	0.36	0.34	1.78	0.18	1.10 [0.96-1.26]	0.07
rs6441091*	chr3:156373580	*TIPARP*	T	C	0.36	0.34	1.90	0.17	1.10 [0.96-1.26]	0.07
rs344066*	chr3:156440305	*TIPARP*	A	G	0.04	0.02	9.83	1.72×10^{-3}	1.73 [1.22-2.45]	0.18
rs1430412	chr3:156521121	*TIPARP-LEKR1*	C	T	0.24	0.23	0.68	0.41	1.07 [0.91-1.24]	0.08
rs9847692*	chr3:156531322	*TIPARP-LEKR1*	C	T	0.25	0.23	0.82	0.36	1.07 [0.92-1.25]	0.08
rs4857612	chr3:177306621	*LINC00578*	G	C	0.39	0.40	0.27	0.60	0.96 [0.84-1.10]	0.07
rs3931397*	chr4:149079497	*NR3C2*	G	T	0.08	0.08	0.37	0.55	1.08 [0.85-1.37]	0.12
rs17024437	chr4:149081808	*NR3C2*	G	A	0.08	0.08	0.37	0.55	1.08 [0.85-1.37]	0.12
rs1309531	chr5:64306311	*CWC27*	A	T	0.43	0.44	1.18	0.28	0.93 [0.81-1.06]	0.07
rs10471310	chr5:64548961	*ADAMTS6*	C	T	0.39	0.38	0.29	0.59	1.04 [0.91-1.19]	0.07
rs10064391	chr5:64686659	*ADAMTS6*	A	G	0.36	0.38	2.14	0.14	0.90 [0.79-1.04]	0.07
rs2047063*	chr5:64732237	*ADAMTS6*	T	C	0.39	0.42	2.21	0.14	0.90 [0.79-1.03]	0.07
rs11746802*	chr5:178665185	*ADAMTS2*	C	T	0.36	0.33	2.70	0.10	1.12 [0.98-1.29]	0.07
rs11743204	chr5:178671014	*ADAMTS2*	T	C	0.38	0.35	2.62	0.11	1.12 [0.98-1.28]	0.07

SNP	Position	Locus	Ref	Alt	Case Freq.	Control Freq.	X^2	P	OR [95% CI]	SE
rs340124	chr5:178686590	*ADAMTS2*	G	A	0.45	0.45	0.03	0.87	0.99 [0.87-1.13]	0.07
rs13191376*	chr6:45522139	*RUNX2*	C	T	0.34	0.36	1.30	0.26	0.92 [0.80-1.06]	0.07
rs1412710	chr6:75837203	*COL12A1*	C	T	0.13	0.16	5.52	0.02	0.79 [0.65-0.96]	0.10
rs1931656	chr6:82610188	*FAM46A*	T	A	0.46	0.45	0.06	0.81	1.02 [0.89-1.16]	0.07
rs9344230*	chr6:82616216	*FAM46A*	T	C	0.30	0.29	0.56	0.46	1.06 [0.92-1.22]	0.07
rs9455877	chr6:169556637	*THBS2*	A	G	0.17	0.15	5.85	0.02	1.24 [1.04-1.48]	0.09
rs11768292*	chr7:65474801	*GUSB*	G	T	0.40	0.42	2.10	0.15	0.91 [0.79-1.04]	0.07
rs3764903*	chr7:66098482	*KCTD7*	G	A	0.47	0.49	0.44	0.51	0.96 [0.84-1.09]	0.07
rs3800817	chr7:66263550	*RABGEF1*	T	A	0.28	0.25	3.99	0.05	1.16 [1.00-1.35]	0.08
rs4717328*	chr7:66352665	*SBDS*	T	C	0.28	0.25	3.47	0.06	1.15 [0.99-1.30]	0.08
rs2106166	chr7:92668332	*SAMD9*	A	T	0.44	0.42	1.17	0.28	1.08 [0.94-1.23]	0.07
rs3808520	chr8:23164773	*LOXL2*	G	C	0.20	0.21	0.88	0.35	0.92 [0.78-1.09]	0.08
rs7026684	chr9:4215308	*GLIS3*	G	A	0.38	0.38	0.02	0.89	0.99 [0.86-1.13]	0.07
rs66720556	**chr9:13559717**	***MPDZ-NFIB***	**T**	**A**	**0.23**	**0.18**	**15.85**	**6.87×10^{-5}**	**1.38 [1.18-1.62]**	**0.08**
rs9409911	chr9:137434446	*RXRA-COL5A1*	A	G	0.30	0.35	10.77	1.03×10^{-3}	0.79 [0.68-0.91]	0.07
rs1536482*	**chr9:137440528**	***RXRA-COL5A1***	**G**	**A**	**0.42**	**0.34**	**24.55**	**7.24×10^{-7}**	**1.40 [1.23-1.61]**	**0.07**
rs3132303	**chr9:137444298**	***RXRA-COL5A1***	**G**	**C**	**0.20**	**0.26**	**17.81**	**2.44×10^{-5}**	**0.70 [0.60-0.83]**	**0.08**
rs11145951*	chr9:139860264	*PTGDS*	C	T	0.45	0.48	3.53	0.06	0.88 [0.77-1.01]	0.07
rs2386136	chr9:139864341	*PTGDS*	G	A	0.45	0.49	3.84	0.05	0.88 [0.77-1.00]	0.07
rs35809595	chr10:63831928	*ARID5B*	G	A	0.44	0.42	1.65	0.20	1.10 [0.96-1.25]	0.07
rs2419835	chr10:115296564	*HABP2*	T	C	0.13	0.13	3.67×10^{-3}	0.95	1.01 [0.83-1.22]	0.10
rs4938174*	chr11:110913240	*C11orf53*	G	A	0.29	0.30	0.12	0.73	0.97 [0.84-1.13]	0.07
rs2242312*	chr11:130275346	*ADAMTS8*	G	A	0.06	0.06	0.07	0.79	0.96 [0.74-1.26]	0.14
rs10859105*	chr12:91473005	*KERA-LUM*	C	T	0.32	0.28	5.48	0.02	1.19 [1.03-1.37]	0.07

SNP	Position	Locus	Ref	Alt	Case Freq.	Control Freq.	X^2	P	OR [95% CI]	SE
rs2268578*	**chr12:91501198**	***LUM***	**G**	**A**	**0.16**	**0.12**	**13.51**	**2.38×10^{-4}**	**1.41 [1.17-1.69]**	**0.09**
rs10859110*	chr12:91504845	*LUM*	G	A	0.29	0.24	10.83	9.97×10^{-4}	1.28 [1.11-1.48]	0.08
rs7308752	chr12:91527181	*DCN*	A	G	0.11	0.08	9.11	2.54×10^{-3}	1.39 [1.12-1.73]	0.11
rs116878472	chr12:104210992	*NT5DC3*	T	C	0.037	0.03	1.52	0.22	1.25 [0.88-1.78]	0.18
rs11553764	chr12:104415244	*GLT8D2*	C	T	0.17	0.18	0.01	0.91	0.99 [0.83-1.18]	0.09
rs2755238	**chr13:41110270**	***FOXO1***	**T**	**C**	**0.15**	**0.10**	**23.92**	**1.00×10^{-6}**	**1.61 [1.33-1.96]**	**0.10**
rs56223983	chr14:81814754	*STON2*	G	T	0.32	0.32	0.03	0.87	0.99 [0.86-1.14]	0.07
rs62014489	chr15:30171879	*TJP1*	G	A	0.10	0.11	1.33	0.25	0.88 [0.71-1.09]	0.11
rs785424	chr15:30178544	*TJP1*	A	T	0.09	0.10	1.38	0.24	0.87 [0.70-1.09]	0.11
rs8030753	chr15:48801935	*FBN1*	C	T	0.14	0.14	0.06	0.80	1.02 [0.85-1.24]	0.10
rs4601989*	chr15:67451954	*SMAD3*	C	T	0.18	0.22	7.94	4.84×10^{-3}	0.78 [0.66-0.93]	0.09
rs12912010	chr15:67467143	*SMAD3*	G	T	0.17	0.22	10.29	1.34×10^{-3}	0.75 [0.63-0.90]	0.09
rs6496932*	chr15:85825567	*AKAP13*	C	A	0.21	0.18	5.85	0.02	1.22 [1.04-1.44]	0.08
rs4843040	chr15:85838636	*AKAP13*	C	T	0.26	0.23	4.65	0.03	1.18 [1.02-1.38]	0.08
rs7183651*	chr15:85895721	*AKAP13*	G	A	0.26	0.23	6.45	0.01	1.22 [1.05-1.42]	0.08
rs7183764*	chr15:85903051	*AKAP13*	G	A	0.26	0.22	5.84	0.02	1.21 [1.04-1.40]	0.08
rs4842882*	chr15:85984183	*AKAP13*	A	G	0.23	0.19	8.12	4.37×10^{-3}	1.26 [1.07-1.47]	0.08
rs2654583*	chr15:101002255	*CERS3*	G	A	0.30	0.30	4.26×10^{-4}	0.98	1.00 [0.86-1.15]	0.07
rs930847*	chr15:101558562	*LRRK1*	T	G	0.20	0.22	2.72	0.10	0.87 [0.74-1.03]	0.08
rs752092*	chr15:101781934	*CHSY1*	A	G	0.33	0.33	1.39×10^{-3}	0.97	1.00 [0.87-1.15]	0.07
rs35193497	chr16:88324821	*BANP-ZNF469*	G	T	0.30	0.34	5.05	0.02	0.85 [0.74-0.98]	0.07
rs8059298	chr16:88332479	*BANP-ZNF469*	C	T	0.32	0.36	5.28	0.02	0.85 [0.74-0.98]	0.07
rs11656734	chr17:7426695	*POL2A*	G	C	0.36	0.39	5.21	0.02	0.85 [0.74-0.98]	0.07
rs4792535*	chr17:14565130	*HS3ST3B1-PMP22*	C	T	0.29	0.29	0.06	0.80	1.02 [0.88-1.18]	0.07

SNP	Position	Locus	Ref	Alt	Case Freq.	Control Freq.	X^2	P	OR [95% CI]	SE
rs9981981*	chr21:47544838	*COL6A2*	G	A	0.04	0.04	0.15	0.70	1.07 [0.76-1.50]	0.17
rs71313932	chr22:19960198	*ARVCF*	G	C	0.29	0.29	4.48×10^{-7}	> 0.99	1.00 [0.86-1.16]	0.07

Ref = the reference allele.
Alt = the alternate allele.
Case Freq. = the frequency of the alternate allele in the case cohort.
Control Freq. = the frequency of the alternate allele in the control cohort.
X^2 = the chi square statistic.
P = the asymptotic p-value.
OR = the estimated odds ratio for the minor allele with the major allele as the reference, where 95% CI indicates the 95% confidence interval for this value.
* Indicates SNPs with hard-typed genotype data available for both cases and controls.
SNPs significantly associated with keratoconus ($p < 6.94 \times 10^{-4}$) are **bold**.

5.5 AIM 2: FINE-MAPPING KERATOCONUS-ASSOCIATED LOCI IN KERATOCONUS CASES AND CONTROLS

5.5.1 Methods

5.5.1.1 Loci selection

Keratoconus loci reaching genome-wide significance ($p < 5 \times 10^{-8}$) in the literature, as well as, any CCT-associated SNPs that were significantly associated with keratoconus following correction for multiple testing in Aim 1 (as described in Section 5.4) were fine-mapped. For each of these loci, the regions surrounding the reported SNP were selected for fine-mapping such that the intergenic region and flanking genes were included for intergenic SNP, whereas the region encompassing the relevant gene was included for intronic SNPs. These broad regions were included to ensure the association peaks were captured.

5.5.1.2 Study participants and genotyping data

The case cohort consisted of 487 unrelated Australian keratoconus patients (described in Section 2.1.1) with genome-wide genotyping data generated on the Illumina HumanCoreExome array (HumanCoreExome-24v1-1_A). The control cohort consisted of 626 controls, 427 were unaffected individuals from the Blue Mountain Eye Study (BMES; as described in Section 2.1.4) and 199 were unscreened population controls from the NSA cohort (described in Section 2.1.5). The controls were genotyped using a customized Illumina HumanCoreExome array ("HumanCoreExome_Goncalo_15038949_A"), as described previously.[254] Only SNPs that were common to both arrays were included in the current analysis.

5.5.1.3 Quality control, imputation and statistical analysis

Quality control and imputation procedures were performed by Dr Bennet McComish (Menzies Institute for Medical Research, University of Tasmania, TAS, Australia). Quality control was conducted according to a modified version of the protocol outlined by Anderson *et al.* (2010).[192] Snpflip (https://github.com/biocore-ntnu/snpflip) was used to detect reverse and ambiguous strand SNPs. Using PLINK, ambiguous strand SNPs were removed from analysis and reverse strand SNPs were flipped. Individuals with a missing genotype rate > 0.08, discordant sex information, or heterozygosity more than three standard deviations from the mean were excluded. Related individuals were detected across both cases and controls by calculating pairwise identity-by-descent (IBD) in PLINK, and the individual with the lower genotyping rate in any pair with IBD > 0.185 was removed from analysis. Ancestry outliers were identified by principal component analysis (PCA) using EIGENSTRAT[193] with the HapMap[191] Phase III reference panel and individuals with a PC1 > 0.07 were excluded. SNPs were excluded if they had a missing genotype rate > 3%, a minor allele

frequency < 0.01, deviated significantly ($p < 10^{-5}$) from Hardy-Weinberg equilibrium, or if the missing call counts differed significantly between cases and controls ($p < 10^{-5}$) as determined by PLINK's case/control non-random missingness test.

Autosomal genotype data was phased using Eagle[255] (version 2.3.5) and genotypes were imputed based on the 1000 Genomes Project reference panel[250] (Phase III, version 5) using Minimac3[252] (version 2.0.1). Using BCFtools (https://github.com/samtools/BCFtools; version 1.3.1), insertions and deletions (indels) were excluded from statistical analysis and SNPs were excluded if they were rare (MAF < 0.01), within 5 bp of an indel, or had poor imputation quality ($r^2 < 0.8$). The resulting VCF file was converted to PLINK format files and multi-allelic SNPs were removed from analysis.

For each locus selected for fine-mapping, the included region was extracted from the PLINK format files. Association analysis was performed for each included region using the most-likely genotypes under a chi-squared model in PLINK as described previously in Section 5.4.1.3.

5.5.1.4 Selecting regions for re-sequencing

Loci with strong association peaks identified in the fine-mapping of keratoconus-associated loci were selected for re-sequencing. At each locus separately, the region under the association peak was selected to include as many of the SNPs with a nominally significant p-value (< 0.05) as were practical. The re-sequenced regions were also influenced by the best sequencing design.

5.5.1.5 Data Visualisation

The results of the association analysis were graphed using the online batch mode of LocusZoom[256] (available at http://locuszoom.org/genform.php?type=hitspecdata). Hard-typed SNPs were plotted as squares, while imputed SNPs were plotted as circles. To identify the most significant SNPs at each locus, the top SNPs were coloured purple and labelled with their variant ID (rsID). All other included SNPs were coloured according to the degree of LD (r^2) with the top SNP. LocusZoom calculated LD estimates with PLINK (version 1.07) using the 'European population hg19/1000 Genomes Nov 2014' build. Regions selected for re-sequencing were highlighted in grey along the lower section of the plot that contains the location of transcripts.

5.5.2 Results

Six loci were selected for fine-mapping, including two novel loci identified in Aim 1 (Section 5.4) and four previously published loci that were associated with keratoconus at a genome-wide significance threshold (Table 5.3). Following quality control procedures, genotyping data from 487 cases and 626 controls were used to fine-map the included regions. Cohort demographics are summarised in Table 5.4.

Table 5.3 – Summary of the keratoconus-associated loci for fine-mapping.

Locus	Associated SNP	P-value	Study	Included region (size)
RAB3GAP1	rs4954218	9.3×10^{-9}	Li *et al.* (2012)[85]; Bae *et al.* (2013)[86]	chr2:135722061-136288806 (5.7 Mb)
FNDC3B	rs4894535	4.9×10^{-9}	Lu *et al.*(2012)[87]	chr3:171757418-172118492 (0.4 Mb)
MPDZ-NFIB	rs1324183	5.0×10^{-8}	Lu *et al.*(2012)[87]; Sahebjada *et al.* (2013)[88]	chr9:3105703-14398982 (11.3 Mb)
RXRA-COL5A1	rs1536482	7.2×10^{-7}	Aim 1 (Section 5.4)	chr9:137208944-137736688 (0.5 Mb)
LUM	rs2268578	2.9×10^{-4}	Aim 1 (Section 5.4); Iglesias *et al.* (2018)[249]	chr12: 91300000-91700000 (0.4 Mb)
FOXO1	rs2721051	2.7×10^{-10}	Lu *et al.*(2012)[87]	13:41048131-41240734 (0.2 Mb)

Associated SNP = the SNP previously associated with keratoconus, either in the literature or from Stage 1 of this study.
P-value = The reported p-value for the associated SNP.

Table 5.4 – Cohort demographics

Group	n	Mean age (range)	% Female	Disease status
Cases	487	44.7 (14-85)	45.0	affected
Controls	199	75.5 (42-96)	53.3	unscreened
	427	60.2 (50-89)	46.2	unaffected

n = the number of individuals in each cohort.
Mean age = the mean average (and the range) is reported in years.

5.5.2.1 Fine-mapping results for the *RAB3GAP1* locus

Fine-mapping of the *RAB3GAP1* locus showed a poor association peak in the present cohort, with no SNPs in high LD with the top SNP, rs4954218 (Figure 5.2). The top SNP obtained a p-value of 2.33 $x10^{-3}$ and is the same SNP that was previously reached genome-wide significance with keratoconus. The minor allele (G) at the top SNP for the *RAB3GAP1* locus, rs4954218, was protective for keratoconus (OR = 0.75). Due to the lack of a strong association peak, the *RAB3GAP1* locus was not selected for re-sequencing in Aim 3.

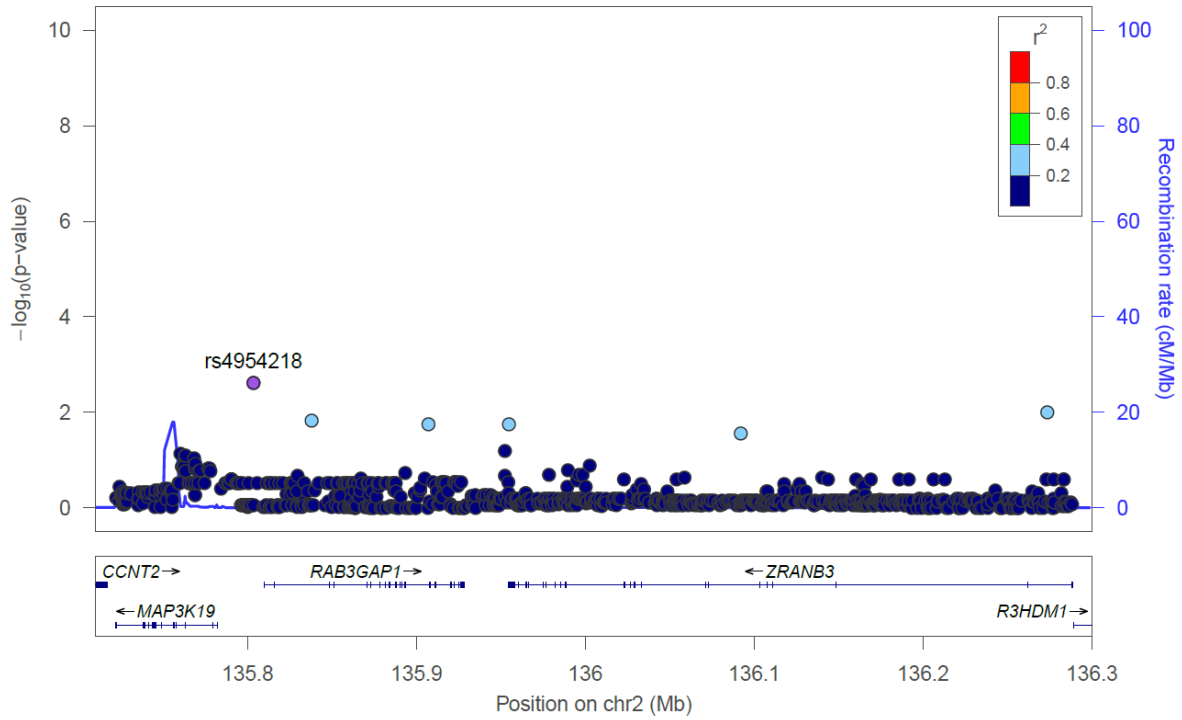


Figure 5.2 – Fine-mapping for the *RAB3GAP1* locus.
Circles indicate imputed SNP data and squares indicate hard-typed SNPs. The top SNP is coloured purple and labelled. All other SNPs are coloured according to their degree of linkage disequilibrium (r^2) with the top SNP based on the European population of the 1000 Genomes (Nov 2014 build). The lower section of the plot indicates the location of nearby genes.

5.5.2.2 Fine-mapping results for the *FNDC3B* locus

For the *FNDC3B* locus, a strong association peak was identified with the top SNP, rs4894538, reaching a p-value of 2.44 $x10^{-3}$ (Figure 5.3). The top SNP in the present study, rs4894538, is 3400 bp downstream of the published keratoconus-associated SNP at this locus, rs4894535. Based on the fine-mapping, a 107 kbp region encompassing the association peak was selected for re-sequencing.

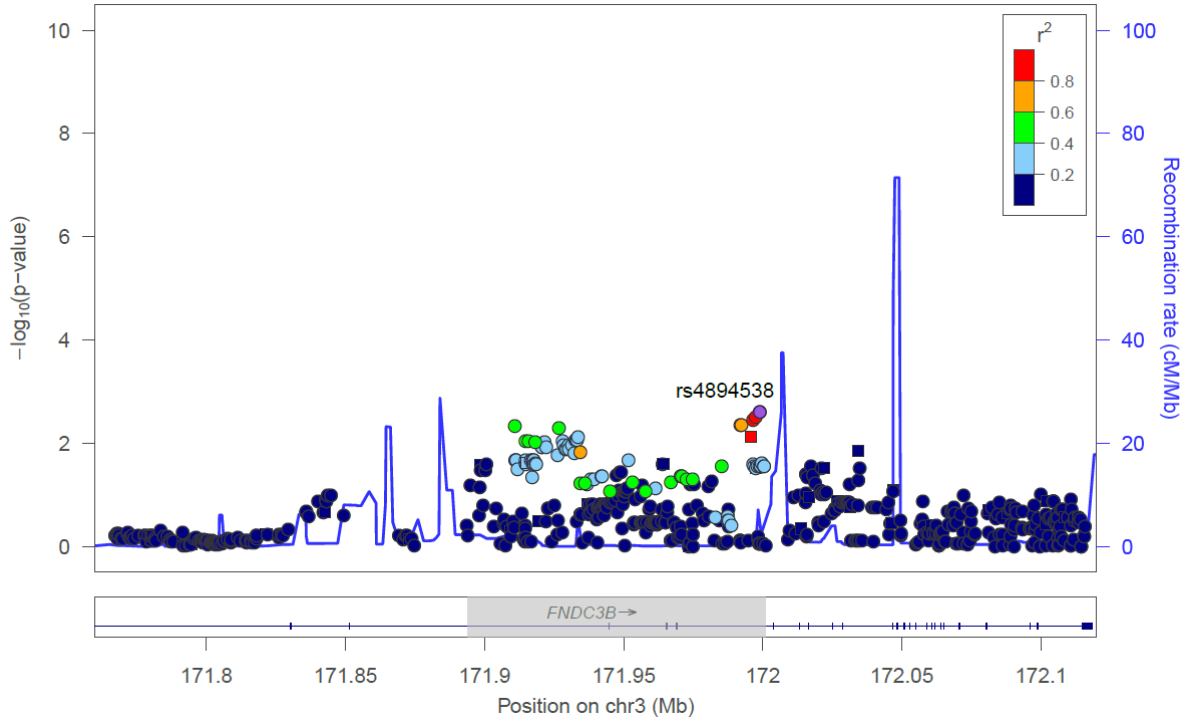


Figure 5.3 – Fine-mapping for the *FNDC3B* locus.
Circles indicate imputed SNP data and squares indicate hard-typed SNPs. The top SNP is coloured purple and labelled. All other SNPs are coloured according to their degree of linkage disequilibrium (r^2) with the top SNP based on the European population of the 1000 Genomes (Nov 2014 build). The lower section of the plot indicates the location of the gene. The grey section indicates the region selected for re-sequencing.

5.5.2.3 Fine-mapping results for the *MPDZ-NFIB* locus

A tight association peak was identified at the *MPDZ-NFIB* locus (Figure 5.4). The top SNP, rs7851770, obtained a p-value of 8.24 x10^{-5} and is located 256 bp away from the previously reported keratoconus-associated SNP, rs1324183. A 36 kbp region under the peak was selected for re-sequencing.

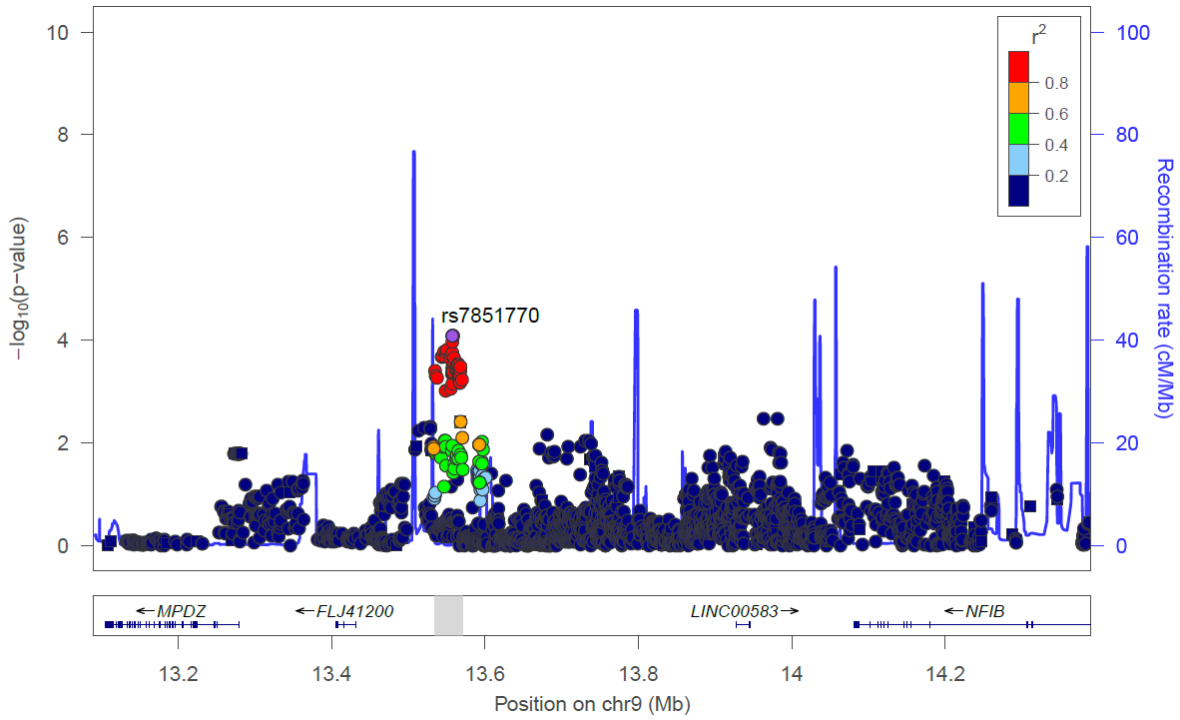


Figure 5.4 – Fine-mapping for the *MPDZ-NFIB* locus.
Circles indicate imputed SNP data and squares indicate hard-typed SNPs. The top SNP is coloured purple and labelled. All other SNPs are coloured according to their degree of linkage disequilibrium (r^2) with the top SNP based on the European population of the 1000 Genomes (Nov 2014 build). The lower section of the plot indicates the location of nearby genes. The grey section indicates the region selected for re-sequencing.

5.5.2.4 Fine-mapping results for the *RXRA-COL5A1* locus

Following fine-mapping, a very tight association peak was identified at the *RXRA-COL5A1* locus (Figure 5.5). The top SNP, rs1536483, reached a p-value of 1.45 x10^{-5}, which was the smallest p-value obtained across all loci that were fine-mapped. The SNP rs1536483 is 156 bp away from the keratoconus-associated SNP rs1536482, which implicated this locus in keratoconus susceptibility in Section 5.4. As few SNPs were captured immediately downstream of the rs1536483, this region was selected for re-sequencing in addition to the region under the association peak.

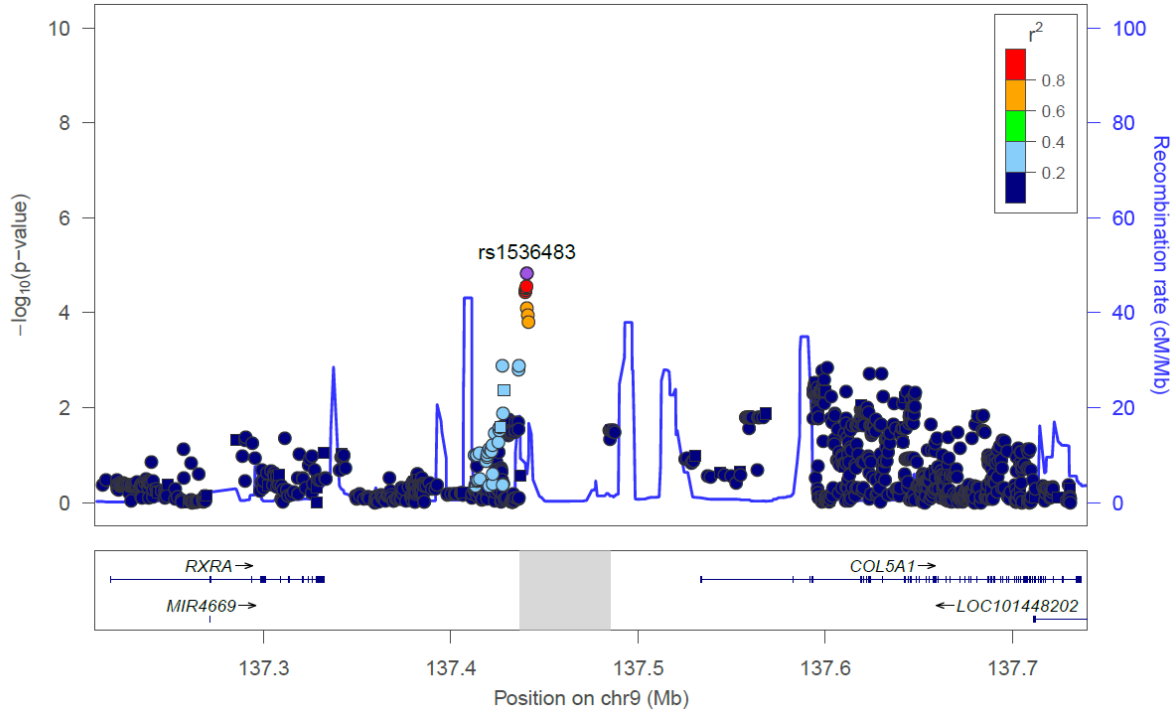


Figure 5.5 – Fine-mapping for the *RXRA-COL5A1* locus.
Circles indicate imputed SNP data and squares indicate hard-typed SNPs. The top SNP is coloured purple and labelled. All other SNPs are coloured according to their degree of linkage disequilibrium (r^2) with the top SNP based on the European population of the 1000 Genomes (Nov 2014 build). The lower section of the plot indicates the location of the nearby genes. The grey section indicates the region selected for re-sequencing.

5.5.2.5 Fine-mapping results for the *LUM* locus

The association peak at the novel *LUM* locus reached a minimum p-value of 4.51 x10^{-3} at rs3759221 (Figure 5.6). Apart from a small spike surrounding the top SNP, the association peak for this locus was very broad and therefore a 131 kbp region, encompassing the genes *KERA*, *LUM* and the majority of *DCN*, was selected for resequencing.

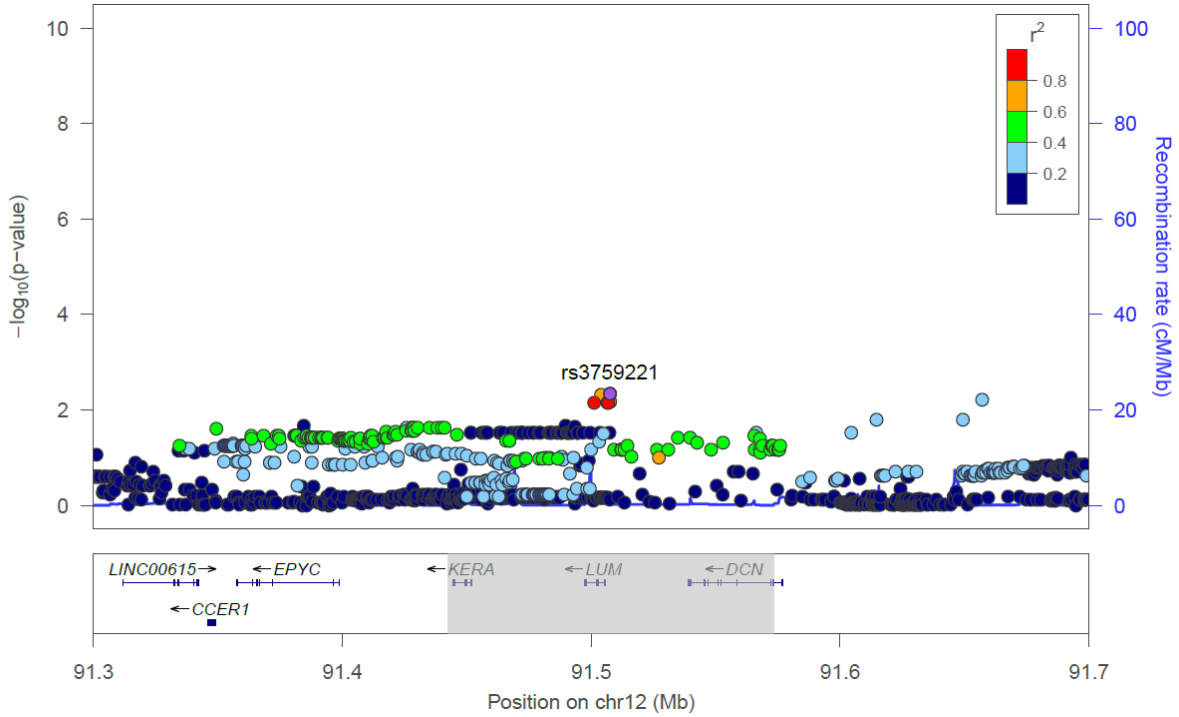


Figure 5.6 – Fine-mapping for the *LUM* locus.
Circles indicate imputed SNP data and squares indicate hard-typed SNPs. The top SNP is coloured purple and labelled. All other SNPs are coloured according to their degree of linkage disequilibrium (r^2) with the top SNP based on the European population of the 1000 Genomes (Nov 2014 build). The lower section of the plot indicates the location of nearby genes. The grey section indicates the region selected for re-sequencing.

5.5.2.6 Fine-mapping results for the *FOXO1* locus

Two apparently independent association peaks were identified at the *FOXO1* locus as the top SNP, rs2755209, and the second top SNP, rs79728429, with a pairwise r^2 value below 0.2 in the European population of the Nov 2014 build of 1000 Genomes (Figure 5.7). In the present cohort however, all homozygotes for the risk-allele (T) at rs79728429 (the rarer of the two SNPs) were also homozygotes for the risk-allele (C) at rs2755209. Moreover, all heterozygotes for the risk-associated allele at rs79728429 were either heterozygous or homozygous for the risk-associated allele at rs2755209. It was therefore hypothesised that both risk-alleles are on the same haplotype and the low r^2 value is indicative of the difference in allele frequency. Following fine-mapping, a 28.7 kbp region underneath the association peak was selected for re-sequencing.

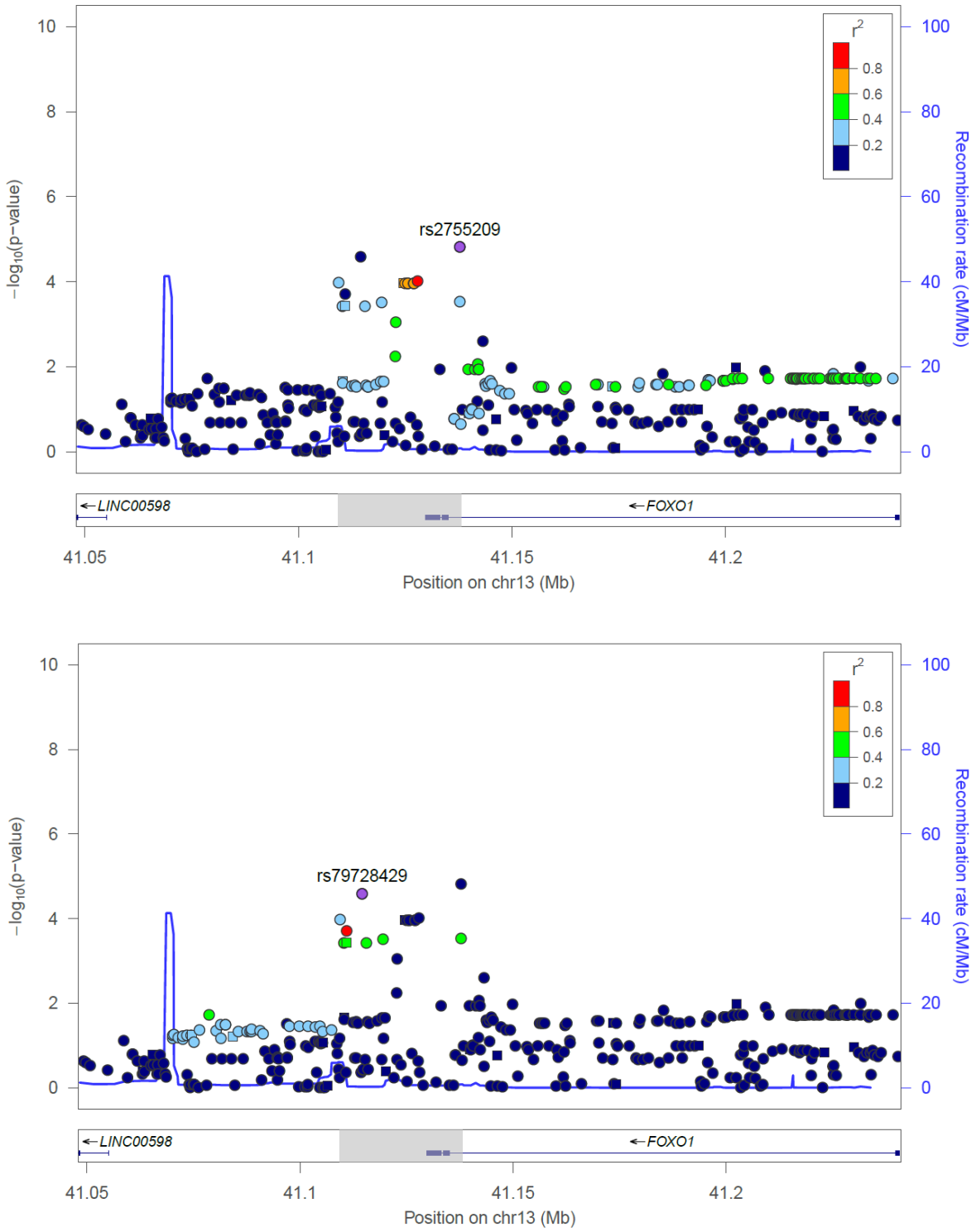


Figure 5.7 – Fine-mapping for the *FOXO1* locus.
Circles indicate imputed SNP data and squares indicate hard-typed SNPs. The two top SNPs are not in linkage disequilibrium (LD), therefore, the top plot displays the top SNP in purple with all other SNPs coloured according to their degree of LD (r^2) with the top SNP. The bottom plot displays the second top SNP in purple and all other SNPs are coloured according to their degree of LD with the second top SNP based on the European population of the 1000 Genomes (Nov 2014 build). The lower section of the plot indicates the location nearby genes. The grey section indicates the region selected for re-sequencing.

5.5.2.7 Summary of the fine-mapping results

The top SNPs at each locus following fine mapping are summarised in Table 5.5. Apart from the *RAB3GAP1* locus, the minor allele at top SNP(s) for all loci conferred an increased risk of keratoconus, with odds ratios above 1. For *RAB3GAP1*, the minor allele at the top SNP was associated with a decreased risk of keratoconus, and therefore is protective. However, due to the limited association peak, the *RAB3GAP1* locus was not selected for re-sequencing. For the remaining loci, the regions selected for re-sequencing are also presented in Table 5.5.

Table 5.5 – A summary of the top SNPs at each locus and the regions following fine-mapping.

Locus	Top SNP	Major Allele	Minor Allele	Case Freq.	Control Freq.	P-value	OR [95% CI]	Regions selected for re-sequencing
RAB3GAP1	rs4954218	T	G	0.27	0.33	2.33×10^{-3}	0.75 [0.63-0.90]	not selected
FNDC3B	rs4894538	A	T	0.24	0.18	2.44×10^{-3}	1.37 [1.12-1.69]	chr3:171893935-172001141
MPDZ-NFIB	rs7851770	G	T	0.23	0.16	8.24×10^{-5}	1.53 [1.24-1.88]	chr9:13534659-13570728
RXRA-COL5A1	rs1536483	C	T	0.42	0.33	1.45×10^{-5}	1.47 [1.23-1.74]	chr9:137436967-137484947
LUM	rs3759221	A	G	0.17	0.12	4.51×10^{-3}	1.41 [1.11-1.79]	chr12:91442305-91573359
FOXO1	rs2755209	A	C	0.43	0.34	1.50×10^{-5}	1.46 [1.23-1.74]	chr13:41109288-41138020
	rs79728429	C	T	0.10	0.05	2.54×10^{-5}	1.99 [1.44-2.75]	

Top SNP = the top SNP for the locus following fine-mapping.
Case freq. = the frequency of the minor allele in the case cohort.
Control freq. = the frequency of the minor allele in the control cohort.
OR [95% CI] = the odds ratio for the minor allele with the major allele as reference and the 95% confidence interval.

5.6 AIM 3: RE-SEQUENCING OF KERATOCONUS-ASSOCIATED LOCI IN CASES AND CONTROLS

5.6.1 Methods

5.6.1.1 Selecting regions for re-sequencing

Five regions were selected for re-sequencing following the fine-mapping of key keratoconus-associated loci as described in Aim 2 (Section 5.5): *FNDC3B*, MPDZ-*NFIB*, *RXRA-COL5A1*, *KERA-LUM-DCN* and *FOXO1*.

5.6.1.1 Study participants

The case cohort consisted of 178 unrelated keratoconus cases, previously described in Section 2.1.1. For each locus, keratoconus patients carrying the risk allele at the top SNPs identified in the fine-mapping analysis in Section 5.5 were selected for re-sequencing. To predict the carrier status, PLINK was used to extract the genotypes at the top SNPs from the fine-mapping data (Section 5.5.1.3). For each locus, individuals that were homozygous carriers of the risk allele were prioritised over heterozygotes to ensure that any variants identified were carried on the risk-associated haplotype. The control cohort consisted of 62 screened individuals without keratoconus from the Blue Mountains Eye Study. This cohort is described in Section 2.1.4. Unlike the cases, the controls were not selected based on their genotypes at the top SNPs, but instead on DNA availability.

5.6.1.2 Re-sequencing

A custom Nextera Rapid Capture Enrichment (Illumina) was designed for the regions of interest following the fine-mapping of keratoconus-associated loci using Illumina's DesignStudio and hg19 as the reference. The target regions are summarised in Table 5.6.

Table 5.6 – A summary of the target regions for the re-sequencing capture.

Locus	Region	Size (bp)	Probes
FNDC3B	chr3:171893935-172001141	107,206	467
MPDZ-NFIB	chr9:13534659-13570728	36,069	157
RXRA-COL5A1	chr9:137436967-137484947	47,980	209
LUM	chr12:91442305-91573359	131,054	570
FOXO1	chr13:41109288-41138020	28,732	125

Probes = indicates the number of probes designed to cover the region selected for re-sequencing.

Both the capture and re-sequencing was conducted in-house. DNA samples were uniquely barcoded, multiplexed (12-plex), and enriched libraries were generated as outlined in Illumina's 'Nextera Rapid Capture Enrichment Reference Guide' (document #15037436 v01; January 2016). Throughout this protocol, a TapeStation (Agilent Technologies, Santa Clara, California, USA) was used to assess the distribution of each library using either Agilent Technologies's d1000 or High Sensitivity d1000 reagents as required. DNA concentrations were determined using a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA). Following enrichment, the nanomolar (nM) concentration for each pooled library was determined as follows:

$$\text{concentration in nM} = \frac{(\text{concentration in ng/ul})}{(660\text{g/mol} \times \text{average fragment size})} \times 10^6$$

Enriched libraries were pooled at equimolar concentrations such that 48 DNA samples were included in each sequencing run. These pooled libraries were diluted to 15 pM, spiked with 1% 20 pM PhiX (1.33% v/v) and denatured with NaOH as outlined in Illumina's "MiSeq System: Denature and Dilute Guide" (document #15039740 v01; January 2016). Paired-end read sequencing was performed on an Illumina MiSeq using 150V3 reagents.

5.6.1.3 Sequence data analysis

Using bcbio-nextgen (https://github.com/bcbio/bcbio-nextgen), FASTQ files were aligned to the reference genome, GRCh37, using BWA[188] (version 0.7.17) and variants were joint-called with GATK[125] (version 3.8) in line with GATK's best practices. This generated in a single variant call format (VCF) file. To obtain locus-specific VCF files, BCFtools (https://github.com/samtools/BCFtools; version 1.5) was used to extract each re-sequenced region separately from the single VCF file. In the same command, multi-allelic variants were split into separate entries and indels were normalised and left aligned. Confidence tags were added to the genotypes as described in Section 2.2.1. Low confidence genotypes at the top SNPs (or second top SNP where relevant) were converted to missing variant calls as described in Section 2.2.2.

To investigate coverage across the re-sequencing regions, depth information at each base-pair position was extracted for all individuals from the aligned BAM files using SAMtools[164] (version 1.8). An example command is outlined in Appendix 16. For each locus, the mean depth and standard deviation was calculated across all samples at all base-pair positions and plotted using the ggplot2[172] package in R.[126] An example of the R script used to calculate the coverage statistics and generate the plot are presented in Appendix 17.

5.6.1.4 Variant annotation

Variants were annotated as outlined in Section 2.2.3. Key annotations used in the present study included variant identification codes (IDs) from the dbSNP[128] 147 database, gene annotations from the

RefGene[128] database, minor allele frequencies (MAF) from the Genome Aggregation Database[129] (gnomAD), as well as, deleteriousness/pathogenicity predictions for SNPs with Combined Annotation Dependent Depletion[134] (CADD) and Functional Analysis through Hidden Markov Models[135] (FATHMM), using the FATHMM-MKL[136] algorithm. Pathogenicity and deleteriousness predictions for small insertions or deletions were annotated separately via the online batch submissions available for CADD (https://cadd.gs.washington.edu/score) and the FATHMM-indel[137] algorithm (http://indels.biocompute.org.uk/) and were manually added to the annotated file. Using R, the frequency of each variant in the cases and controls, as well as maximum MAF observed across the eight ethnic populations in gnomAD (African, Admixed American, Ashkenazi Jewish, East Asian, Finnish, Non-Finnish European and 'Other', which includes individuals without an assigned population), were calculated and added to the annotated file.

For each locus separately, pairwise linkage disequilibrium (LD) correlations, D' and r^2, between the top SNP and all other variants identified at the locus were determined using HaploView.[257] To ensure the variant IDs were unique this column in the locus-specific VCFs were updated to a concatenation of the chromosome number, position, reference and alternate alleles (in the format: 1:1234C,T) using R. The VCF files with the updated ID columns were then converted to binary PLINK format files in PLINK whilst using the 'keep-allele-order' option to ensure the reference and alternate alleles were coded appropriately. To ensure HaploView could handle the insertions and deletions, these files were converted to PLINK format MAP and PED files using the modifier '--recode 12' to code reference alleles as '1's and alternate alleles as '2's. Phenotype information was also added to the PED file at this stage. To match the input format for HaploView, a new file was generated from the PLINK MAP file, excluding the columns containing the chromosome and centimorgan. This file, along with the unaltered PED file were uploaded to the HaploView under the linkage format option. In HaploView, all markers were selected in the 'check markers tab' and the data from the 'LD plot' tab was downloaded as a text file. This file contained pairwise LD correlations between all possible pairs of variants in the given locus, namely D' and r^2 values. Comparisons between the top SNP (or second top SNP if relevant) and all other variants were transferred into a separate text file. These data were combined with the annotated file using the merge function in R.

5.6.1.5 Variant prioritisation

To ensure that low frequency variants carried on the risk-associated haplotype were included during variant prioritisation, D' was used to measure pairwise LD correlation between the top SNP and all other SNPs identified within each locus. Along with the D' value, HaploView outputs a log of the likelihood odds ratio (LOD) which represents the confidence in the D' value. This LOD score is defined as log10(L1/L0), where L1 is likelihood of the data under linkage disequilibrium and L0 is likelihood of the data under linkage equilibrium. Together, the D' and LOD scores were used to identify variants

that were in high LD with the top SNP with high confidence. For each locus separately, variants were filtered to identify those with a D' greater than 0.9; a LOD greater than 2; and variants that obtained high confidence genotype calls (depth ≥ 10 and quality ≥ 20) in at least 50% of the re-sequenced individuals, including multiple high confidence variant calls. Variants were further prioritised to include only variants that were present in more than one case; observed at a frequency more than 1% higher in the cases compared to the control cohort; and observed at a frequency more than 1% higher in the cases compared to the maximum frequency observed in the gnomAD populations (or absent in gnomAD).

5.6.1.6 Identifying putatively functional variants

Highly prioritised variants were further investigated to identify putatively functional variants as previously described in Section 2.2.4. CADD and FATHMM scores were also considered at this stage of analysis.

5.6.2 Results

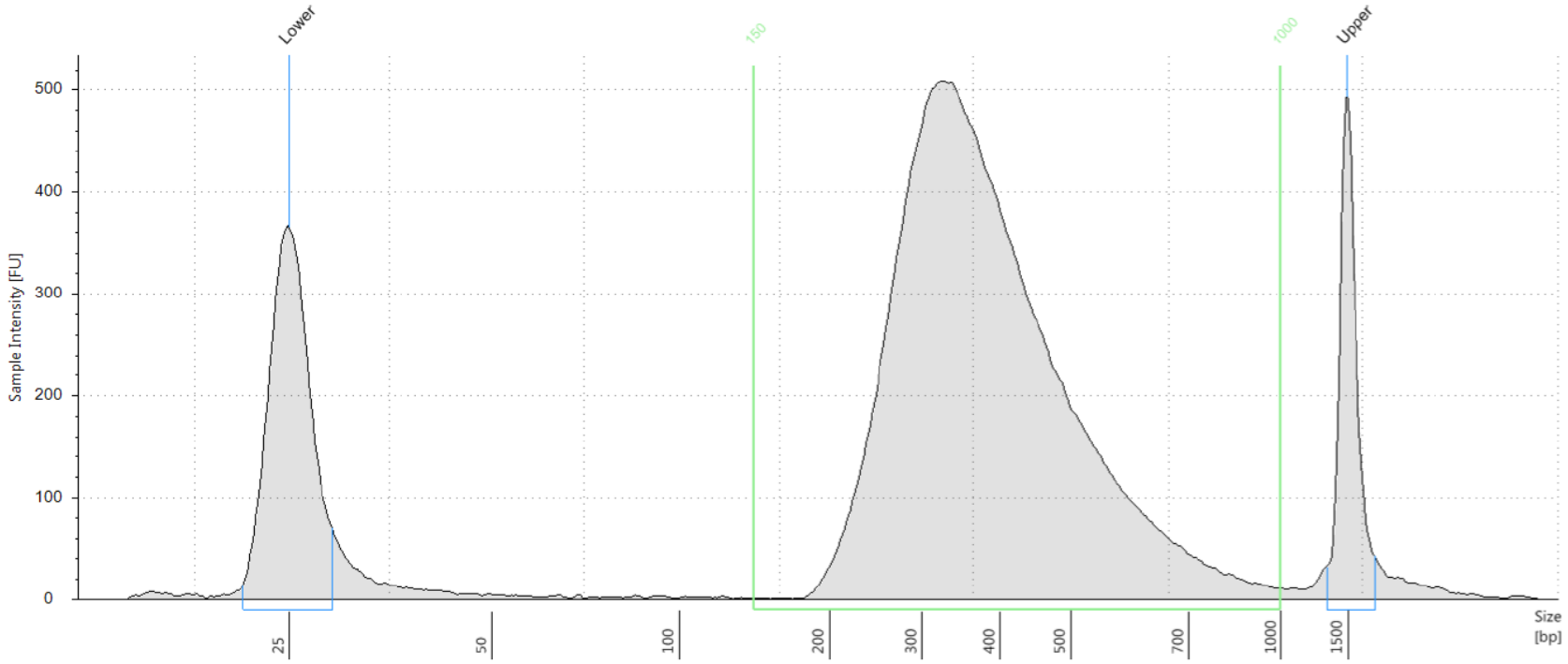
The five keratoconus-associated regions with strong association peaks in the fine-mapping analysis were re-sequenced in 178 keratoconus cases and 62 population controls. The cohort demographics are summarised in Table 5.7.

Table 5.7 – Cohort demographics

Group	n	Mean age (range)	% Female	Disease status
Cases	178	45.8 (16-85)	44.4	affected
Controls	62	65.1 (51-73)	48.4	unaffected

n = the number of individuals in each cohort.
Mean age = the mean average (and the range) is reported in years.

Enriched libraries (containing DNA from 12 individuals) were of high quality with a mean fragment size of 384.8 (368-402) for fragments between 150 bp and 1000 bp. A representative electropherogram is presented in Figure 5.8.



Region Table

From [bp]	To [bp]	Average Size [bp]	Conc. [pg/µl]	Region Molarity [pmol/l]	% of Total	Region Comment	Color
150	1000	394	2290	9750	94.45		

Figure 5.8 – A representive electropherogram of an enriched library analysed with Agilent's 4200 TapeStation system using high sensitivity d1000 tapes.

The electropherogram demonstrates the size distribution of the DNA library, where the x-axis indicates the size of the DNA fragments and the y-axis indicates the sample intensity. The lower and upper markers are labelled and gated in blue. The region between 150 bp and 1000 bp is gated in green and represents the DNA library. Details for this region are outlined in the region table below the plot.

Re-sequencing for all 240 individuals was conducted across five sequencing runs, with 48 individuals included in each run. Sequencing quality metrics are outlined in Table 5.8.

Table 5.8 – Re-sequencing metrics by sequencing run

Run	Cluster Density	Clusters PF (%)	Total Reads	Reads PF	Total % ≥ Q30
1	1,439 ±25	86.55 ±1.46	34,096,024	29,517,874	93.01
2	1,518 ±25	79.99 ±1.71	35,089,808	28,070,918	91.90
3	1,656 ±38	80.65 ±1.75	37,810,504	30,510,564	90.13
4	1,498 ±34	84.97 ±1.09	35,109,640	29,839,604	89.25
5	1,656 ±31	74.30 ±3.64	36,546,592	27,201,368	85.51

PF = "passing filters".
% ≥ Q30 = the percentage of reads with a quality score greater or equal to 30.

Sufficient coverage (depth ≥ 10) for variant calling was obtained for 84.9% to 96.4% of the target region, depending on the locus (Table 5.9).

Table 5.9 – Coverage statistics by locus

Locus	Total bases	Bases with depth < 10 (n regions)	% with depth < 10
FNDC3B	107,207	8,418 (79)	7.85
MPDZ-NFIB	36,070	2,465 (13)	6.83
RXRA-COL5A1	47,981	7,254 (36)	15.12
LUM	131,055	4,675 (51)	3.57
FOXO1	28,733	2,751 (21)	9.57

Total bases = the total bases in the re-sequenced target region.
Bases with depth < 10 = the number of bases with a mean depth less than 10.
n regions = the number of regions with a mean depth less 10, where a region may be an isolated base with depth below 10 or multiple consecutive bases with depth below threshold.
% with depth < 10 = the percentage of the re-sequenced target region with a mean depth less than 10

The re-sequencing data was used to determine the carrier status that the top SNP for each locus as the fine-mapping data was imputed and the genotypes at these SNPs were not known in the control cohort. These data are outlined in Table 5.10. For the *FNDC3B* locus, the top SNP from the fine-mapping

analysis (rs4894538) was found to be within a repetitive sequence and therefore the genotypes at this SNP were poorly called. For this reason, the second top SNP (rs4894414), located 1,506 bp downstream of rs4894538, was used for all subsequent analysis of this locus. For the *FOXO1* locus, three individuals had low confidence genotype calls for the top SNP (rs2755209), and therefore the genotype for these individuals were converted to missing so they did not affect the pairwise LD estimates with the other identified variants. A comparison of the top SNPs selected for the re-sequencing analysis and the previously reported SNPs are presented in Table 5.11.

Table 5.10 – Carrier status at the SNPs of interest for each re-sequenced region.

				Cases				Controls				Total			
Locus	**Top SNP**	**Ref**	**Alt**	**Hom**	**Het**	**NC**	**Miss**	**Hom**	**Het**	**NC**	**Miss**	**Hom**	**Het**	**NC**	**Miss**
FNDC3B	rs4894414*	C	T	28	60	90	0	4	17	41	0	32	77	131	0
MPDZ-NFIB	rs7851770	G	T	23	70	85	0	4	17	41	0	27	87	126	0
RXRA-COL5A1	rs1536483	C	T	54	76	48	0	6	27	29	0	60	103	77	0
LUM	rs3759221	A	G	11	58	109	0	0	16	46	0	11	74	155	0
FOXO1	rs2755209	C	T	51	90	35	2	12	27	22	1	63	117	57	3

Ref = the reference allele.
Alt = the alternate allele.
Hom = the number of homozygotes (individuals carrying two copies of the alternate allele).
Het = the number of heterozygotes (individuals carrying one copy of the alternate allele).
NC = the number of non-carriers (individuals with no copies of the alternate allele).
Miss = the number of individuals with missing (unknown) genotype calls.
*Indicates the second top SNP for the locus and that this SNP was used for analysis.

Table 5.11 – A comparison of the top SNPs selected for the re-sequencing analysis and the previously reported SNPs

Locus	Reported SNP	Top SNP	Ref	Alt	Alternate allele frequencies			D' (LOD)	r^2	Distance (bp)
					Cases	Controls	gnomAD max			
FNDC3B	rs4894535	rs4894414*	C	T	0.3119	0.1855	0.3207	1.00 (89.1)	0.93	1,894
MPDZ-NFIB	rs1324183	rs7851770	G	T	0.6629	0.7903	0.8544	1.00 (91.2)	0.95	256
RXRA-COL5A1	rs1536482	rs1536483	C	T	0.5028	0.3145	0.4708	1.00 (102.8)	0.96	156
LUM	rs2268578	rs3759221	A	G	0.7865	0.9113	0.9109	0.93 (51.5)	0.76	6,363
FOXO1	rs2721051	rs2755209	C	T	0.2247	0.0887	0.0925	0.97 (17.1)	0.21	26,920

Reported SNP = The keratoconus-associated SNP reported in the literature or from Aim 1 of the present study.
Top SNP = The SNP with the smallest p-value identified at the locus in the fine-mapping experiment (Aim 2).
Ref = the reference allele.
Alt = the alternate allele.
gnomAD max = the maximum frequency of alternate allele observed across the populations available in the gnomAD database.
D' (LOD) = the pairwise D prime value for the reported SNP and the top SNP. The LOD score for the D prime value is presented in the parenthesis.
r^2 = the pairwise r squared value for the reported SNP and the top SNP.
Distance (bp) = the distance between the reported SNP and the top SNP in base pairs.
*Indicates the second top SNP for the locus and that this SNP was used for analysis in Aim 3.

5.6.2.1 Re-sequencing results for the *FNDC3B* locus

Coverage across the target region at the *FNDC3B* locus was generally of high quality with mean depth of 140 and standard deviation (sd) of 37.3. A total of 8,418 bases spread across 79 regions had insufficient read depth for high confidence variant calling, corresponding to 7.85% of the re-sequenced region. The largest of these was a region 1.9 kb region (3:171959424-171961339) that co-located in a long terminal repeat element as observed on the ReapeatMasker[258] track available on the UCSC Genome Browser. A coverage plot for this locus is presented in Figure 5.9.

A total of 853 variants were identified across the re-sequenced region. As highlighted previously, the second top SNP (rs4894414) at this locus was used as a proxy for the top SNP. Five variants were highly prioritised, including four SNPs and one deletion (Table 5.12). Despite meeting all other filtering conditions, the reported SNP for the *FNDC3B* locus from the literature (rs4894535) was not included in this list as the alternate allele was more frequent in the East Asian population in gnomAD (32.1%) than our keratoconus cases (31.2%). The highly prioritised variants were all located within introns of *FNDC3B* and span approximately a 50 kb region. None of these variants obtained CADD or FATHMM scores indicative of functionality, however, some noteworthy annotations were identified following investigations using the UCSC Genome Browser. Three of the highly prioritised variants, rs76047624, rs77351096 and rs7635832, were located in DNase1 hypersensitivity clusters in six to ten of the cell types available from ENCODE. These three variants, along with rs35417004, are also located within an annotated enhancer region in at least one cell line. Additionally, rs77351096 is located within chromatin immunoprecipitation (ChIP) peaks for 19 transcription factors, indicating that these factors bind within this region.

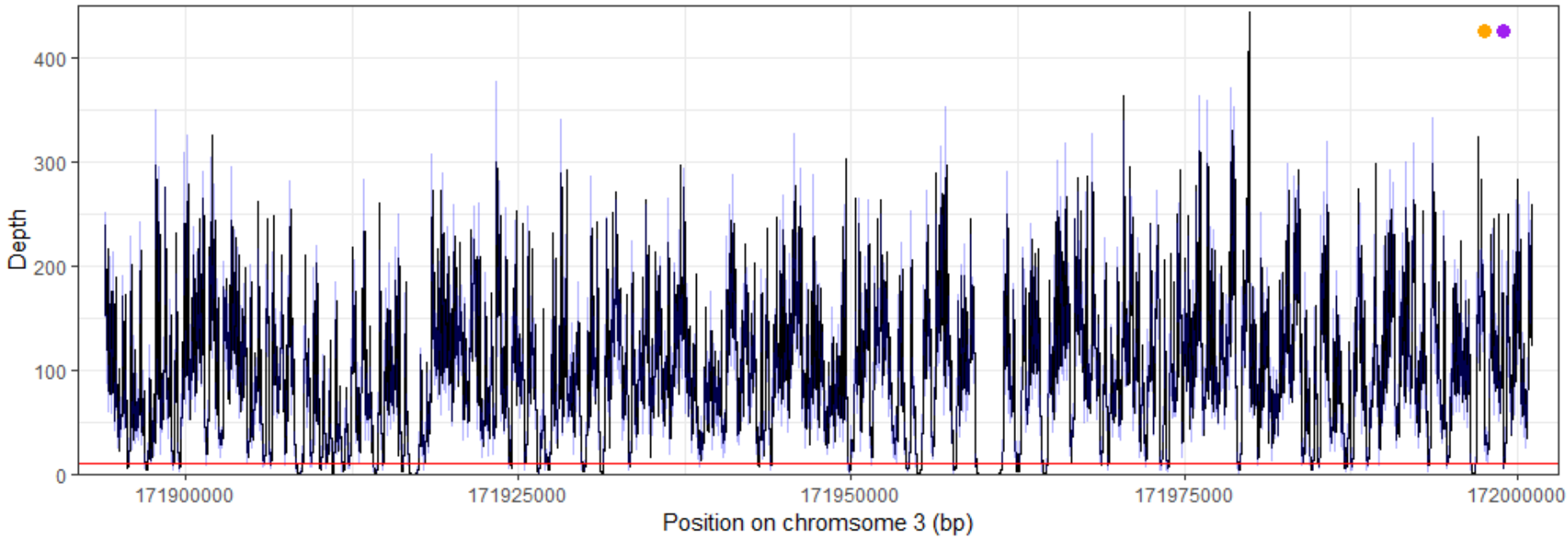


Figure 5.9 – Coverage of the re-sequenced region at the *FNDC3B* locus.
The black line indicates the mean depth, the red horizontal line indicates a depth of 10 (the minimum depth required for variant calling) and the blue shaded area indicates the area between the mean ± 1 standard deviation. The position of the top SNP and second top SNP are indicated by the purple and orange dots (respectively).

Table 5.12 – Highly prioritised variants at the *FNDC3B* locus.

						Alternate allele frequencies				
Position	**Ref**	**Alt**	**Variant ID**	**CADD**	**FATHMM**	**Cases**	**Controls**	**gnomAD max**	**D' (LOD)**	**r^2**
3:171940636-171940636	G	A	rs76047624	7.71	0.19	0.0646	0.0161	0.0542	1.00 (3.30)	0.02
3:171947244-171947244	T	-	rs145117275	3.62	0.03	0.0702	0.0484	0.0336	1.00 (11.93)	0.17
3:171961459-171961459	C	T	rs35417004	4.88	0.06	0.1854	0.1210	0.1097	0.92 (31.14)	0.42
3:171988707-171988707	T	C	rs77351096	1.69	0.18	0.0843	0.0645	0.0504	1.00 (15.38)	0.21
3:171989276-171989276	T	G	rs7635832	1.13	0.16	0.3343	0.2097	0.3240	1.00 (95.69)	0.96
3: 171997499- 171997499	C	T	rs4894414	5.04	0.20	0.3258	0.2016	0.3220		

Ref = the reference allele.
Alt = the alternate allele.
Variant ID = the variant identifier from the avsnp147 database.
CADD = the scaled CADD score.
FATHMM = the FATHHMM-MKL or FATHMM-indel score.
gnomAD max = the maximum frequency of alternate allele observed across the populations available in the gnomAD database.
D' (LOD) = the pairwise D prime value for this variant and the top SNP. The LOD score for the D prime value is presented in the parenthesis.
r^2 = the pairwise r squared value for the variant and the top SNP.
The row containing the information for the top SNP is shaded grey.

5.6.2.2 Re-sequencing results for the *MPDZ-NFIB* locus

A mean depth of 102.6 reads (sd = 29.9) was obtained for the re-sequenced region at the *MPDZ-NFIB* locus. A total of 2,465 bases across thirteen regions had insufficient coverage for variant calling (6.8% of the re-sequenced region) which largely corresponded to short interspersed nuclear elements, visualised using the RepeatMasker track on UCSC Genome Browser. Coverage for the re-sequenced region for *MPDZ-NFIB* is presented in Figure 5.10.

A total of 403 variants were called across the *MPDZ-NFIB* locus. Thirty-nine variants in high LD with risk allele at the top SNP (rs7851770) were highly prioritised following variant filtering, including three deletions and 36 SNPs (Table 5.13). Two SNPs, rs12003602 and rs34074476, obtained CADD scores above 10, which is suggestive of functionality. The FATHMM scores for the same variants confirm this potential for rs12003602 (0.75). This variant removes a CpG site, however further investigation of this variant using the tracks available on the UCSC Genome Browser were unremarkable.

The reported SNP from the literature at the *MPDZ-NFIB* locus (rs1324183) was in high LD with the top SNP identified in the fine-mapping experiment (rs7851770) with a D' of 1 (LOD = 91.2), however, the alternate allele was observed less frequently in our keratoconus patients (66.3%) compared to our controls (79.0%). The frequency of this variant in the control group is consistent with the allele frequency reported in the non-Finnish European population of gnomAD (80.5%). Therefore, this variant was not included in the highly prioritised list of variants at this locus.

Almost all of the highly prioritised variants at this locus were located within regions annotated as heterochromatin in all nine cell lines included in the chromatin state segmentation from ENCODE/Broad, however, seven SNPs overlap an enhancer in at least one of the cell lines: rs11999938, rs34813744, rs35638627, rs35641278, rs1324186, rs35846464 and rs34759288. Six of these are located within a 3 kb window at the 3' end of the re-sequenced region, surrounding a region with strong evidence of a regulatory region (Figure 5.11). Five of these variants are also located within DNase1 hypersensitivity clusters identified in at least seven cell types. Based on these findings, it is likely that the functional variation at the *MPDZ-NFIB* locus is located within this 3kb region.

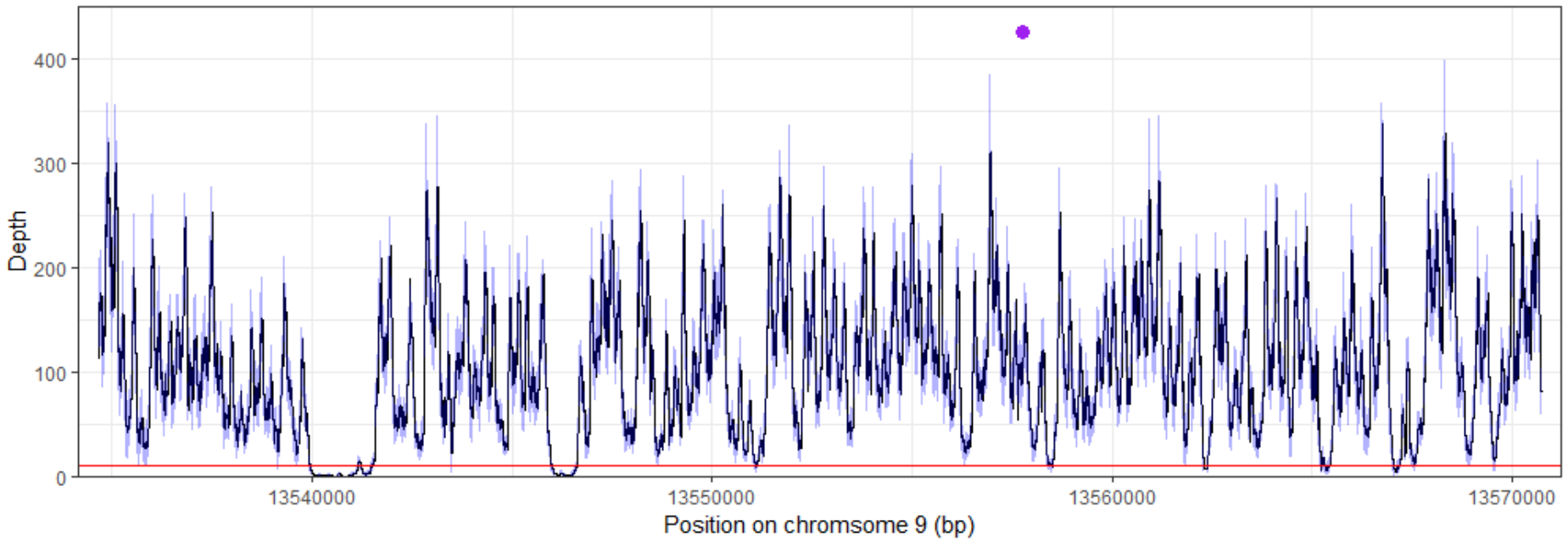


Figure 5.10 – Coverage of the re-sequenced region at the *MPDZ-NFIB* locus.
The black line indicates the mean depth, the red horizontal line indicates a depth of 10 (the minimum depth required for variant calling) and the blue shaded area indicates the area between the mean ± 1 standard deviation. The position of the top SNP is indicated by the purple dot.

Table 5.13 – Highly prioritised variants at the *MPDZ-NFIB* locus.

Position	Ref	Alt	Variant ID	CADD	FATHMM	Alternate allele frequencies			D' (LOD)	r^2
						Cases	Controls	gnomAD max		
9:13534850-13534850	G	A	rs71507380	3.22	0.10	0.1545	0.0887	0.0855	0.98 (26.87)	0.98
9:13534994-13534994	A	G	rs36052788	7.57	0.17	0.3090	0.1774	0.2898	0.99 (79.59)	0.99
9:13536083-13536083	G	A	rs12002959	0.80	0.06	0.3090	0.1613	0.2723	0.99 (76.90)	0.99
9:13537768-13537768	G	A	rs12003602	14.84	0.75	0.3006	0.1613	0.1834	0.99 (72.66)	0.99
9:13543094-13543094	C	G	rs13295903	1.28	0.17	0.0393	0.0161	0.0148	1.00 (6.10)	1.00
9:13543646-13543646	C	T	rs13296965	8.21	0.95	0.3118	0.1613	0.2208	1.00 (82.13)	1.00
9:13544779-13544779	A	G	rs67878159	2.54	0.07	0.3146	0.1613	0.2361	0.99 (79.59)	0.99
9:13544878-13544881	CAAA	-	rs67012432	2.26	0.00	0.3062	0.1613	0.2306	0.99 (75.15)	0.99
9:13546790-13546790	C	A	rs12686335	2.80	0.11	0.3118	0.1855	0.1954	1.00 (86.67)	1.00
9:13548060-13548060	C	A	rs10491754	2.06	0.13	0.3118	0.1613	0.2712	1.00 (82.13)	1.00
9:13548561-13548561	A	G	rs77767334	2.66	0.08	0.0590	0.0403	0.0430	0.93 (8.20)	0.93
9:13550069-13550069	C	A	rs2224860	0.59	0.04	0.3174	0.1613	0.2891	0.98 (78.26)	0.98
9:13550541-13550541	A	G	rs977580	3.12	0.11	0.3258	0.2016	0.2971	0.98 (91.35)	0.98
9:13551273-13551273	C	T	rs7855460	6.28	0.16	0.3230	0.2016	0.2378	0.94 (76.6)	0.94
9:13555074-13555074	G	A	rs13294011	0.78	0.16	0.3202	0.1774	0.2962	1.00 (89.59)	1.00
9:13555912-13555912	**G**	**T**	**rs11999938**	**1.23**	**0.16**	**0.3258**	**0.1774**	**0.2114**	**0.98 (85.37)**	**0.98**
9:13557504-13557504	G	T	rs7851523	0.88	0.09	0.3258	0.1774	0.2967	1.00 (92.86)	1.00
9:13557747- 13557747	G	T	rs7851770	10.25	0.19	0.3258	0.2016	0.3112		
9:13558378-13558378	G	A	rs13291445	0.22	0.05	0.3258	0.2016	0.2474	1.00 (99.73)	1.00
9:13558455-13558455	A	G	rs12686184	0.77	0.07	0.3371	0.1935	0.3089	0.96 (79.68)	0.96
9:13559717-13559717	T	A	rs66720556	1.54	0.15	0.3230	0.1774	0.2469	1.00 (91.17)	1.00
9:13559821-13559821	G	T	rs10491756	0.59	0.19	0.3202	0.1774	0.2447	1.00 (89.29)	1.00
9:13561657-13561658	AT	-	rs35928895	5.90	0.00	0.3230	0.1774	0.2002	1.00 (91.17)	1.00

Position	Ref	Alt	Variant ID	CADD	FATHMM	Alternate allele frequencies			D' (LOD)	r^2
						Cases	Controls	gnomAD max		
9:13562146-13562146	A	C	rs12004620	0.31	0.09	0.3287	0.2016	0.1967	0.99 (93.41)	0.99
9:13563513-13563513	G	T	rs66592246	0.80	0.07	0.3174	0.1774	0.1954	1.00 (87.49)	1.00
9:13563660-13563660	G	T	rs13290289	0.67	0.07	0.3258	0.1694	0.2416	1.00 (90.87)	1.00
9:13565084-13565084	A	-	rs35870971	6.04	0.00	0.3230	0.1855	0.1905	0.98 (85.37)	0.98
9:13565222-13565222	G	T	rs34944131	0.08	0.05	0.3202	0.1935	0.1680	0.92 (70.75)	0.92
9:13565696-13565696	C	G	rs60613246	0.46	0.13	0.3258	0.1694	0.1897	0.98 (83.58)	0.98
9:13565965-13565965	G	T	rs34416482	0.25	0.11	0.3258	0.1694	0.1892	0.98 (83.58)	0.98
9:13566456-13566456	G	A	rs34074476	13.02	0.13	0.3258	0.1694	0.1893	0.98 (83.58)	0.98
9:13566852-13566852	C	T	rs61183216	2.06	0.05	0.3258	0.1694	0.2394	0.98 (83.58)	0.98
9:13566987-13566987	G	A	rs12684945	0.71	0.03	0.3258	0.1774	0.2445	0.98 (85.07)	0.98
9:13567440-13567440	G	A	rs12685187	3.37	0.06	0.3230	0.1694	0.2416	0.99 (85.24)	0.99
9:13567822-13567822	**C**	**T**	**rs34813744**	**4.13**	**0.06**	**0.3202**	**0.1774**	**0.1889**	**0.98 (82.47)**	**0.98**
9:13567945-13567945	**A**	**G**	**rs35638627**	**3.90**	**0.18**	**0.3202**	**0.1774**	**0.1897**	**0.98 (82.47)**	**0.98**
9:13567984-13567984	**G**	**A**	**rs35641278**	**0.71**	**0.14**	**0.3202**	**0.1532**	**0.1875**	**0.98 (77.96)**	**0.98**
9:13568063-13568063	C	T	rs1324186	5.60	0.18	0.4045	0.2581	0.3272	0.94 (52.96)	0.94
9:13568184-13568184	**A**	**G**	**rs35846464**	**2.55**	**0.16**	**0.3202**	**0.1774**	**0.1885**	**0.98 (82.47)**	**0.98**
9:13570462-13570462	**G**	**C**	**rs34759288**	**0.01**	**0.14**	**0.3146**	**0.1532**	**0.1934**	**0.98 (75.58)**	**0.98**

Ref = the reference allele.
Alt = the alternate allele.
Variant ID = the variant identifier from the avsnp147 database.
CADD = the scaled CADD score.
FATHMM = the FATHHMM-MKL or FATHMM-indel score.
gnomAD max = the maximum frequency of alternate allele observed across the populations available in the gnomAD database.
D' (LOD) = the pairwise D prime value for this variant and the top SNP. The LOD score for the D prime value is presented in the parenthesis.
r^2 = the pairwise r squared value for the variant and the top SNP.
The row containing the information for the top SNP is shaded grey.
Variants that overlap an enhancer in at least one of the cell lines in the chromatin state segmentation data from ENCODE/Broad are **bold**.

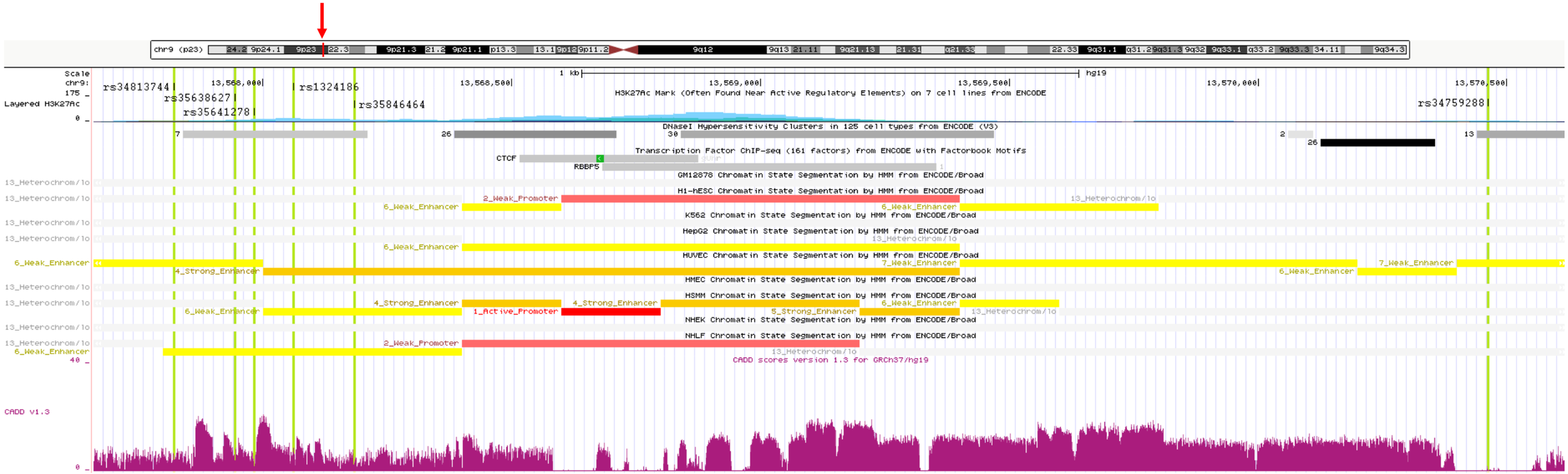


Figure 5.11 – A screenshot from the UCSC Genome Broswer highlighting the location of highly prioritised variants at the *MPDZ* locus surrounding a regulatory region at the distal end of the re-sequenced region.

The red line on the schematic of chromosome 9 (indicated by the red arrow) shows the location of the ~ 3 kb region displayed in the plot (chr9:13,567,661-13,570,621). The green vertical highlights indicate the position of the highly prioritised variants identified. The tracks from top to bottom are: layered H3K27Ac marks on 7 cell lines from ENCODE, where blue peaks indicate regions identified in Human umbilical vein endothelial cells (HUVEC); DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3), where the number to the right of the grey box indicates the number of cell types the cluster is found in; transcription factor ChIP-seq (161 factors) from ENCODE, where the darker the shade of grey the more transcription factors that bind to the region indicated; Chromatin State Segmentation by HMM from ENCODE/Broad ('pack' display mode), where promotors are red, strong enhancers are orange, weak or poised enhancers are yellow, weakly transcribed regions are light green, dark green indicates transcriptional transition or elongation, blue identifies an insulator, dark grey indicates polycomb-repressed DNA and light grey represents heterochromatin, repetitive DNA or copy number variation; and CADD v1.3 where the higher the pink peak, the higher the scaled score.

5.6.2.3 Re-sequencing results for the *RXRA-COL5A1* locus

The re-sequenced region at the *RXRA-COL5A1* locus obtained a mean coverage of 90.6 (sd = 30.6). A coverage plot is presented in Figure 5.12. Thirty-six regions had insufficient coverage for variant calling (7,254 bases in total), corresponding to 15.12% of the re-sequenced region. Many of these regions overlap repetitive elements as observed on the RepeatMasker track on the UCSC Genome Browser. The largest of these regions with insufficient coverage was 2,865 bp in length (chr9:137478643-137481507) and co-located with a long interspersed nuclear element. Conversely, a ~2,200 bp region (9:137464558-137466773) within the re-sequenced region was obtained a mean depth greater than 200 reads. This region overlaps an insulator – a long-range regulatory element that functions by blocking enhancers from acting on promotor regions – as annotated in the chromatin state segmentation track available on the UCSC Genome Browser.

A total of 356 variants were identified across all individuals in the re-sequenced region for the *COL5A1-RXRA* locus. Five SNPs were highly prioritised following variant filtering, including rs1536482, the SNP reported in the literature that originally implicated this locus in keratoconus (Table 5.14). These SNPs are all within a 750 bp window, just upstream of the top SNP identified in the fine-mapping analysis (rs1536483). All five SNPs are located in a region that is weakly transcribed in embryonic stem cells (H1-hESC), human skeletal muscle myoblasts (HSMM) and normal human lung fibroblasts (NHLF) and is repressed or forms heterochromatin in the remaining six cell lines available in the chromatin state segmentation track on the UCSC Genome Browser. The variants also all overlap DNase1 hypersensitivity clusters in at least five of the 125 cells lines, with rs3118517 overlapping clusters in the most cells types (14). While none of the highly prioritised variants obtained CADD or FATHMM scores indicative of functionality, rs3118517 obtained the highest scores for both algorithms with a CADD score of 5.52 and a FATHMM score of 0.10.

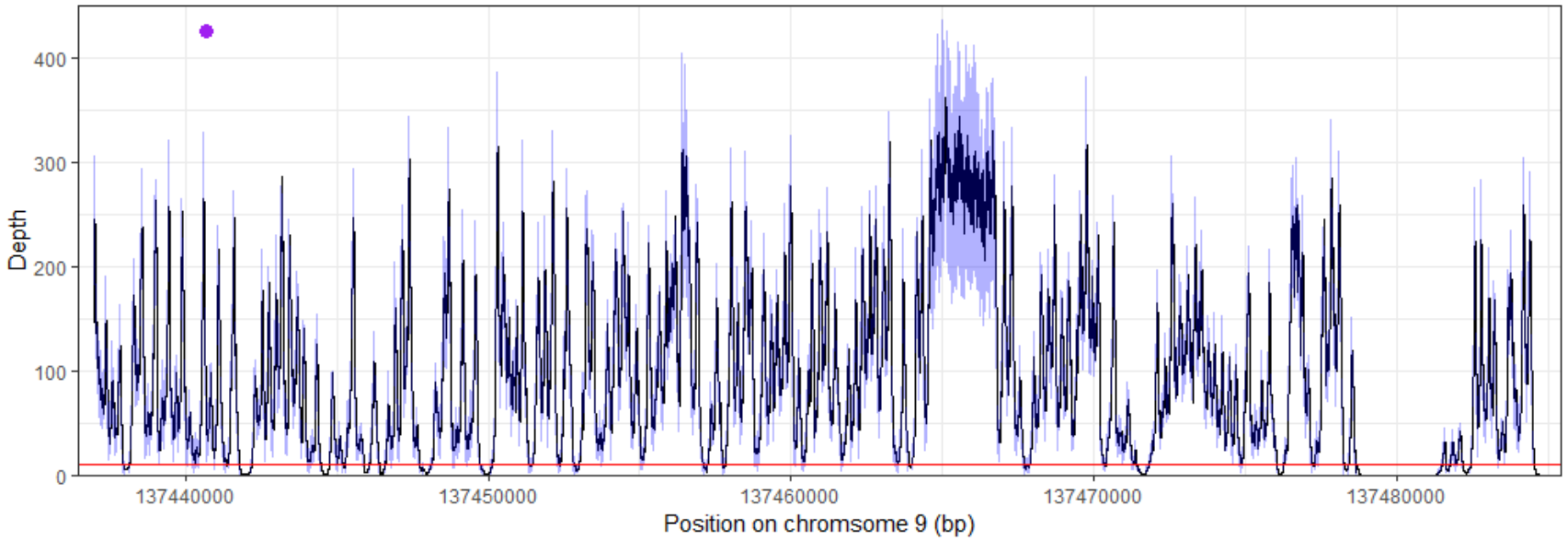


Figure 5.12 – Coverage of the re-sequenced region at the *RXRA-COL5A1* locus.
The black line indicates the mean depth, the red horizontal line indicates a depth of 10 (the minimum depth required for variant calling) and the blue shaded area indicates the area between the mean ± 1 standard deviation. The position of the top SNP is indicated by the purple dot.

Table 5.14 – Highly prioritised variants at the *RXRA-COL5A1* locus.

Position	Ref	Alt	avsnp147	CADD	FATHMM	Alternate allele frequencies			D' (LOD)	r^2
						Cases	Controls	gnomAD max		
9:137439792-137439792	G	A	rs3118516	0.36	0.04	0.5028	0.3065	0.4172	1.00 (101.44)	0.95
9:137439905-137439905	C	T	rs3118517	5.52	0.10	0.5028	0.3065	0.4238	1.00 (101.44)	0.95
9:137440083-137440083	G	A	rs3118518	0.66	0.05	0.5056	0.3065	0.4205	0.99 (98.68)	0.94
9:137440212-137440212	T	C	rs3132306	3.72	0.09	0.5084	0.3065	0.4205	0.98 (96.83)	0.93
9:137440528-137440528	G	A	rs1536482	0.39	0.05	0.5028	0.3145	0.4708	1.00 (102.82)	0.96
9:137440684-137440684	C	T	rs1536483	1.95	0.06	0.5169	0.3145	0.5194		

Ref = the reference allele.
Alt = the alternate allele.
Variant ID = the variant identifier from the avsnp147 database.
CADD = the scaled CADD score.
FATHMM = the FATHHMM-MKL or FATHMM-indel score.
gnomAD max = the maximum frequency of alternate allele observed across the populations available in the gnomAD database.
D' (LOD) = the pairwise D prime value for this variant and the top SNP. The LOD score for the D prime value is presented in the parenthesis.
r^2 = the pairwise r squared value for the variant and the top SNP.
The row containing the information for the top SNP is shaded grey.

5.6.2.1 Re-sequencing results for the *KERA-LUM-DCN* locus

A mean depth of 104.6 (sd = 30.3) was obtained for the re-sequenced region at the *KERA-LUM-DCN* locus. A total of 4,675 bases across fifty-one regions had insufficient coverage for high confidence variant calling, corresponding to 3.6% of the re-sequenced region. This locus was the most comprehensively captured locus included in the re-sequencing experiment. A coverage plot is presented in Figure 5.13.

1,039 variants were identified across the re-sequenced region. It is noteworthy that all of the variants identified at this locus were non-coding despite the re-sequencing region encompassing three genes (*KERA*, *LUM* and *DCN*). Ten variants were highly prioritised following variant filtering, including seven SNPs, two insertions and one deletion (Table 5.15). All of these are low frequency variants with maximum frequencies in the gnomAD database below 4%. None of these variants obtained CADD or FATHMM scores indicative of functionality, however, one of the insertions, rs535582722, is located in the 3' untranslated region of decorin (*DCN*) in six transcripts available in NCBI RefSeq genes track on the UCSC Genome Browser (NM_001920.4, and NM_133503.3- NM_133507.3, NM_133504.3, NM_133505.3 NM_133506.3, and NM_133507.3). This variant is also located within a region that is annotated as a weak enhancer identified in two cell lines (HUVEC and NHLF) and overlaps a ChIP peak for a transcription factor, making it the most likely putatively functional variant at this locus. All other variants are located in intronic or intergenic regions. The variants 12:91500467insA and rs538786536 were also located in a weak enhancer region in one of the nine cell lines (NHLF and HSMM, respectively). The 12:91500467insA is absent in the gnomAD database, but is present in both our cases and controls, suggesting that it is either a common, population specific polymorphism or a sequencing artefact.

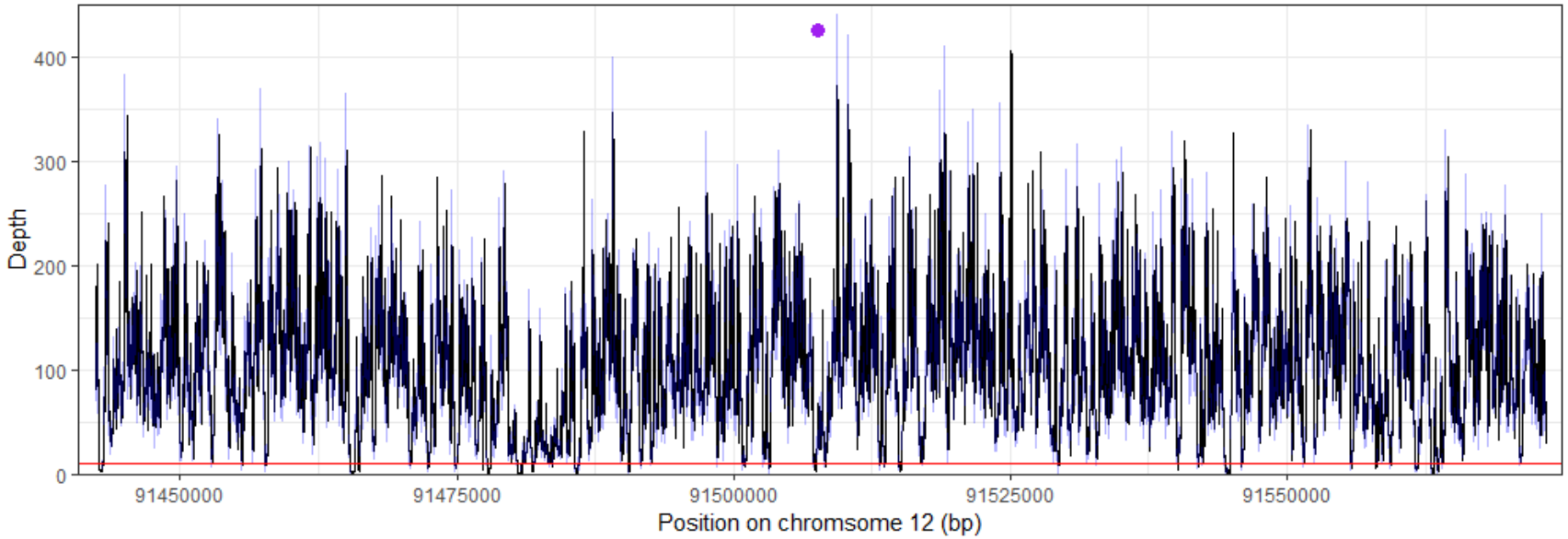


Figure 5.13 – Coverage of the re-sequenced region at the *LUM* locus.
The black line indicates the mean depth, the red horizontal line indicates a depth of 10 (the minimum depth required for variant calling) and the blue shaded area indicates the area between the mean ± 1 standard deviation. The position of the top SNP is indicated by the purple dot.

Table 5.15 – Highly prioritised variants at the *LUM* locus.

Position	Ref	Alt	Location	Variant ID	CADD	FATHMM	Alternate allele frequencies			D' (LOD)	r^2
							Cases	Controls	gnomAD max		
12:91483743-91483743	T	-	*KERA-LUM*	rs776115787	0.80	0.00	0.0140	0.0000	0.0033	1.00 (2.34)	0.04
12:91483745-91483745	T	C	*KERA-LUM*	rs747712265	0.73	0.19	0.0140	0.0000	0.0033	1.00 (2.34)	0.04
12:91491354-91491354	A	C	*KERA-LUM*	rs73197102	1.24	0.07	0.0815	0.0323	0.0375	1.00 (17.84)	0.30
12:91500467-91500467	-	A	*LUM* (intronic)	novel	0.67	0.01	0.1404	0.1129	0.0000	1.00 (2.15)	0.04
12:91507561-91507561	A	G	*LUM-DCN*	rs3759221	2.52	0.03	0.2247	0.1290	0.6409		
12:91515237-91515237	G	A	*LUM-DCN*	rs73198626	0.54	0.03	0.0197	0.0000	0.0082	1.00 (3.48)	0.06
12:91537633-91537633	-	TA	*DCN* (3' UTR)	rs535582722	0.06	0.01	0.0281	0.0081	0.0170	1.00 (4.91)	0.09
12:91541802-91541802	C	G	*DCN* (intronic)	rs143915956	0.40	0.13	0.0281	0.0081	0.0164	1.00 (4.91)	0.09
12:91549814-91549814	A	G	*DCN* (intronic)	rs538786536	0.63	0.15	0.0169	0.0000	0.0066	1.00 (3.06)	0.05
12:91564805-91564805	T	G	*DCN* (intronic)	rs73198634	6.21	0.14	0.0618	0.0000	0.0173	1.00 (11.53)	0.19
12:91566250-91566250	C	T	*DCN* (intronic)	rs3138185	4.53	0.18	0.0787	0.0161	0.0334	1.00 (16.22)	0.27

Ref = the reference allele.
Alt = the alternate allele.
Variant ID = the variant identifier from the avsnp147 database.
CADD = the scaled CADD score.
FATHMM = the FATHHMM-MKL or FATHMM-indel score.
gnomAD max = the maximum frequency of alternate allele observed across the populations available in the gnomAD database.
D' (LOD) = the pairwise D prime value for this variant and the top SNP. The LOD score for the D prime value is presented in the parenthesis.
r^2 = the pairwise r squared value for the variant and the top SNP.
The row containing the information for the top SNP is shaded grey.

5.6.2.2 Re-sequencing results for the *FOXO1* locus

Mean coverage across the re-sequenced region for the *FOXO1* locus was 92.2 (sd = 28.6). A total of 2,751 bases had insufficient coverage for high confidence variant calling, corresponding to 9.57% of the re-sequenced region. A plot of the coverage for the re-sequenced region at *FOXO1* is presented in Figure 5.14.

A total of 198 variants were identified across the *FOXO1* locus. Nine variants were prioritised for further investigation following variant filtering, including eight SNPs and a small insertion (Table 5.16). The reported SNP in the literature for the *FOXO1* locus, rs2721051, was included in this filter. The nine highly prioritised variants largely clustered in the second intron of the ENST00000636651.1 transcript of *FOXO1*, around the single exon gene *AL133318.1* (Figure 5.15).

AL133318.1 (ENSG00000269120.1) is a newly discovered gene located at chr13:41111138-41111323, which only appears to be in the most recent release of the GENCODE Project[259] (V28lift37; April 2018). This gene encodes an uncharacterised protein consisting of 62 amino acids. According to the GETx database, the mRNA for *AL133318.1* is expressed at very low levels in all tissues assessed, with relatively high expression in sun exposed skin, the most relevant tissue to cornea. Within the ~31 kb region presented in Figure 5.15 that encompasses all highly prioritised variants, there are seven expression quantitative trait loci (eQTL) for *AL133318.1*. Of these eQTLs, three were identified in five or six tissues including sun exposed skin (rs58104999, rs7333246 and rs7332960), one is an eQTL in subcutaneous adipose tissue (rs17446593) and three are eQTLs in the basal ganglia region of the brain (rs79703116, rs80223071 and rs1078892). All nine of the highly prioritised variants identified in the present study are within a 1.5 kb window of an eQTL for *AL133318.1* in at least one tissue type and five – 13:41110554insTTTTCTTTC, rs2721051, rs74948688, rs79728429, rs1078892 – are within 250 bp of the eQTLs rs58104999, rs7333246 and rs1078892.

Of the nine highly prioritised SNPs, rs79728429 is most compelling putatively functional variant as it is located in an enhancer region annotated in eight of the nine cell lines avaliable in the chromoatin state segmentation by Hidden Markov Model from ENCODE/Broad and overlaps ChIP seq peaks for 14 transcription factors. The SNP is also located within a strong H3K27Ac histone mark in Human umbilical vein endothelial cells (HUVEC) and overlaps DNase1 Hypersenitivity clusters in 92 of the 125 cell types from ENCODE. This SNP was also the second most associated SNP in the fine-mapping analysis of this locus described in Aim 2. Taken together, these annotations suggest that this variant is located in a likely regulatory element. An eQTL for *AL133318.1*, rs7333246, is also located within this same regulatory region, just 132 bp away from rs79728429, and has been shown to significantly affect the expression of the AL133318.1 protein in five of the tissues in the GETx database, including sun exposed skin.

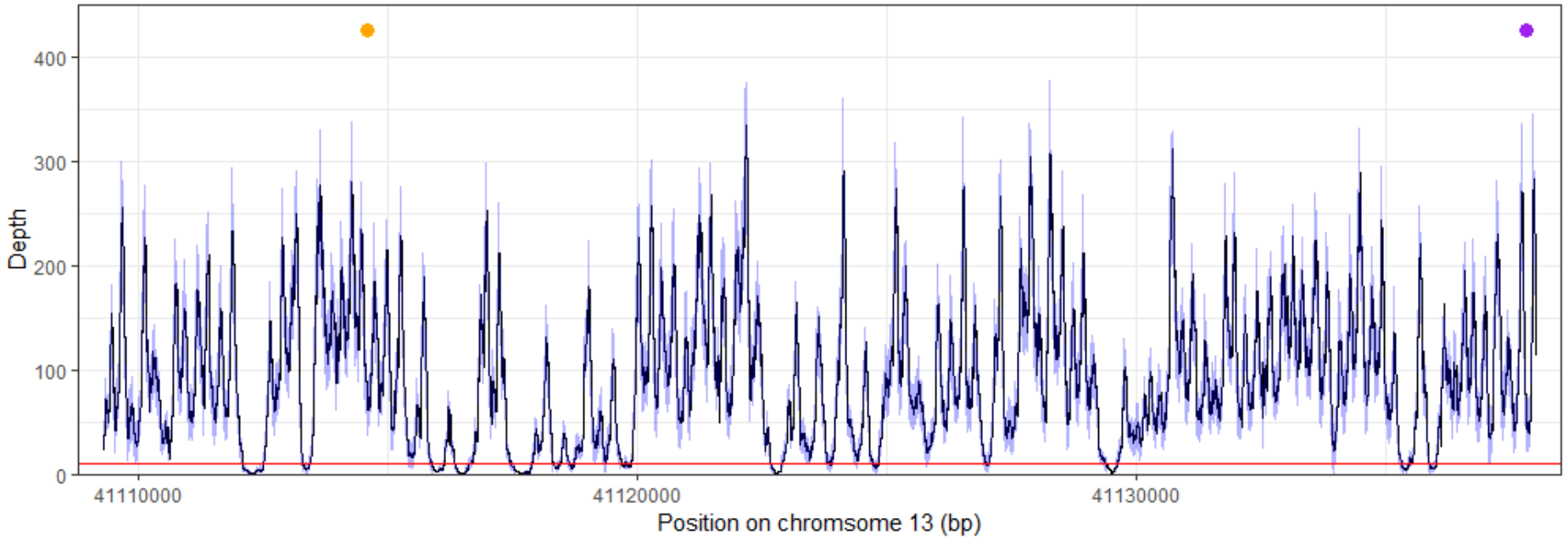


Figure 5.14 – Coverage of the re-sequenced region at the *FOXO1* locus.
The black line indicates the mean depth, the red horizontal line indicates a depth of 10 (the minimum depth required for variant calling) and the blue shaded area indicates the area between the mean ± 1 standard deviation. The position of the top SNP and second top SNP are indicated by the purple and orange dots (respectively).

Table 5.16 – Highly prioritised variants at the *FOXO1* locus.

Position	Ref	Alt	Variant ID	CADD	FATHMM	Alternate allele frequencies			D' (LOD)	r^2
						Cases	Controls	gnomAD max		
13:41110270-41110270	T	C	rs2755238	2.40	0.15	0.2247	0.0887	0.0926	0.97 (17.07)	0.21
13:41110554-41110554	-	TTTTCTTTC	NA	0.09	0.00	0.1910	0.0806	0.0299	0.96 (12.60)	0.18
13:41110884-41110884	C	T	rs2721051	11.62	0.34	0.2247	0.0887	0.0925	0.97 (17.07)	0.21
13:41110922-41110922	C	T	rs74948688	15.43	0.33	0.1713	0.0484	0.0606	0.96 (12.22)	0.14
13:41114572-41114572	C	T	rs79728429	14.84	0.34	0.1573	0.0565	0.0556	0.95 (10.52)	0.13
13:41115586-41115586	A	C	rs2701857	0.17	0.07	0.2247	0.0887	0.0922	0.97 (17.07)	0.21
13:41119466-41119466	G	A	rs11616662	0.08	0.03	0.2247	0.0887	0.1467	0.97 (17.07)	0.21
13:41126936-41126936	C	G	rs2701894	1.76	0.12	0.4916	0.3629	0.4695	0.92 (58.16)	0.69
13:41137784-41137784	G	A	rs80070740	4.66	0.26	0.2247	0.0968	0.0968	1.00 (20.21)	0.23
13:41137804-41137804	A	C	rs2755209	2.65	0.13	0.5393	0.4113	0.7667		

Ref = the reference allele.
Alt = the alternate allele.
Variant ID = the variant identifier from the avsnp147 database.
CADD = the scaled CADD score.
FATHMM = the FATHHMM-MKL or FATHMM-indel score.
gnomAD max = the maximum frequency of alternate allele observed across the populations available in the gnomAD database.
D' (LOD) = the pairwise D prime value for this variant and the top SNP. The LOD score for the D prime value is presented in the parenthesis.
r^2 = the pairwise r squared value for the variant and the top SNP.
NA = the variant did not have a variant ID available in the avsnp147 database**.**
The row containing the information for the top SNP is shaded grey.

Figure 5.15 – A screenshot from the UCSC Genome Broswer highlighting the location of highly prioritised variants at the *FOXO1* locus in high LD with rs2755209 that were identified by re-seqeuncing.
The red line on the schematic of chromosome 13 (indicated by the red arrow) shows the location of the ~31 kb region displayed in the plot (chr13:41,109,745-41,140,649). The green vertical highlight indicates the position of rs79728429, the putatively functional variant at this locus. All other highly prioritised variants are highlighted in yellow. The tracks from top to bottom are: UCSC genes; GENCODE V28lift37 genes, note the small blue square indicating the position of *AL133318.1* on the top left; combined expression QTLs (eQTLs) from 44 tissues from GTEx (midpoint release, V6), where the position of SNPs that are eQTLs for *AL133318.1* in multiple tissues (including sun exposed skin) are indicated in red; SNPedia; layered H3K27Ac marks on 7 cell lines from ENCODE, where blue peaks indicate regions identified in Human umbilical vein endothelial cells (HUVEC); DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3), where the number to the right of the grey box indicates the number of cell types the cluster is found in; transcription factor ChIP-seq (161 factors) from ENCODE, where the darker the shade of grey the more transcription factors that bind to the region indicated; Chromatin State Segmentation by HMM from ENCODE/Broad, where promotors are red, strong enhancers are orange, weak or poised enhancers are yellow, weakly transcribed regions are light green, dark green indicates transcriptional transition or elongation, blue identifies an insulator, dark grey indicates polycomb-repressed DNA and light grey represents heterochromatin, repetitive DNA or copy number variation; and CADD v1.3 where the higher the pink peak, the higher the scaled score.

5.7 DISCUSSION

Aim 1 of this study identified a novel keratoconus-associated locus at rs2268578, located in the first intron of the lumican gene (*LUM*), by assessing CCT-associated loci in our keratoconus cohort. This stage of the study used CCT as an endophenotype for keratoconus, an approach that had previously been shown to be an effective method for identifying keratoconus-associated loci and was integral in the identification of three of the four GWAS hits for keratoconus reported to date.[87] The present study contributed to a large publication that conducted a cross-ancestry GWAS for CCT and subsequently screened novel CCT-associated loci in disease cohorts, including keratoconus.[249] While the CCT GWAS included in the published work informed SNP selection in this dissertation, the published study only assessed the sentinel SNP at each CCT-associated locus in the keratoconus patients (36 SNPs), and a number of SNPs were excluded as they were not available in the other keratoconus cohorts. In the published work, the *LUM* locus was represented by rs7308752, located between *LUM* and *DCN*, and showed a suggestive association ($p = 6.33 \times 10^{-3}$) with keratoconus following a meta-analysis of our cases (n = 711), keratoconus patients from Northern Island (n = 135) and a cohort of keratoconus patients from the USA (n = 240).[249] In this dissertation, this work was extended by fine-mapping the *LUM* locus in our cohort of cases and controls using genome-wide imputation data (Aim 2) which identified a broad association peak that encompassed not only *LUM*, but also keratocan (*KERA*) and decorin (*DCN*). The present study also demonstrated that the top SNP at this locus was rs3759221.

All three genes at the *KERA-LUM-DCN* locus belong to the small leucine-rich proteoglycan family,[260] and together encode three of the four small leucine-rich proteoglycans present in the stroma of the adult cornea.[32, 261] Proteoglycans are macromolecules composed of a protein core with at least one covalently linked glycosaminoglycan chain. In the cornea, proteoglycans play crucial roles in collagen fibril assembly, matrix organization, and ultimately, corneal transparency,[32, 262] however they also perform diverse functions beyond these roles. Decorin binds growth factors allowing it to modulate a number of biological pathways, including TGF-β;[263] plays a role in cell growth and differentiation and inhibits apoptosis;[264] is involved in cutaneous wound healing and angiogenesis;[265] and the expression of *DCN* is known to be regulated by cytokines, including TNF-α,[266] IL-1[267] and IL4-4,[268] suggesting a dynamic role in inflammation. Lumican has similar roles in apoptosis;[269, 270] corneal epithelial wound healing;[271] cell migration and invasion;[270, 272] inflammation;[272] and angiogenesis.[273] There is also evidence that lumican is involved in the expression of *KERA*.[274] While there are fewer studies that examine the function of keratocan, it has been shown to modulate cell differentiation and bone formation,[275] and plays a unique role in the overall development and shape of the cornea with keratocan-null mice displaying narrow iridocorneal angles (the angle between the iris and the cornea), thin stromas with abnormal packaging and organisation of the collagen fibrils, but normal corneal transparency.[262] Clearly, this locus is highly relevant to the normal development, structure and maintenance of the cornea and therefore is a highly plausible keratoconus-susceptibility locus.

Variation within *KERA*, *LUM* and *DCN* have previously been linked to ocular diseases and/or corneal abnormalities, which adds additional evidence for their important roles in the cornea. Truncating variants in *DCN* cause congenital stromal corneal dystrophy (OMIM 610048) which features opaque flecks in the stroma and abnormal organisation of the lamellae, although normal corneal thickness.[276] A number of variants located in *LUM* have been associated with high myopia in studies largely conducted in the Han Chinese population,[277-287] although other studies in the Han Chinese, as well as, Korean and Japanese populations do not support this association.[288-293] Though less relevant to the cornea, *LUM* has also been implicated in the pathogenesis of glaucoma with one study demonstrating a 1.5-fold increase in *LUM* expression trabecular meshwork of individuals with primary open-angle glaucoma compared to healthy controls.[294] *KERA* has previously been linked to two autosomal recessive corneal dystrophies. The first is cornea plana type 2 (OMIM 217300), which is caused by homozygous or compound heterozygous variants in *KERA* and features flattened corneas, corneal opacities and indistinct limbus, resulting high hypermetropia (long-sightedness).[295] Similarly, homozygous or compound heterozygous variants in the gene encoding carbohydrate sulfotransferase 6 (*CHST6*), which encodes the enzyme that mediates the sulfation of keratocan in the cornea, results in macular corneal dystrophy (OMIM 217800).[296] This dystrophy is characterised by progressive punctate opacities of the cornea, which ultimately results in bilateral vison loss.[296] Increased expression of *KERA* has also been observed in the stroma of keratoconic corneas, compared to healthy controls.[297] Interestingly, though the re-sequenced region in Aim 3 of the present study encompassed all three genes, no protein-coding variants were identified in any of the cases or controls. This suggests that the functional variant at this locus is likely to be located in a regulatory region. Following variant filtering, ten variants were highly prioritised at this locus, and while evidence of the functionality of these variants was limited, a low frequency variant (rs535582722) located in the 3' UTR region of *LUM* showed the most promise. Further investigations are required to determine the functional variant and target gene(s) at this novel and compelling keratoconus-susceptibility locus.

Three additional loci were significantly associated with keratoconus in the first stage of the present study including two SNPs between *RXRA* and *COL5A1* – rs1536482 ($p = 7.24 \times 10^{-7}$) and rs3132303 ($p = 2.4 \times 10^{-5}$) – and two known keratoconus-susceptibility loci: *FOXO1* (rs2755238; $p = 1.0 \times 10^{-6}$) and *MPDZ-NFIB* (rs66720556; $p = 6.9 \times 10^{-5}$). While this study lacked power to reach genome-wide significance, it should be highlighted that the most significantly associated locus was *RXRA-COL5A1*, with rs1536482 surpassing the well-known genome-wide associated keratoconus loci. This suggests that variation within the *RXRA-COL5A1* locus is likely to be important in keratoconus susceptibility, which is further supported by the fact that *COL5A1* encodes an alpha chain for collagen type V, a fibril-forming collagen that is highly expressed in the corneal stroma.[298] The significant association we observed in our cohort is in line with previous studies conducted in similar populations. The SNP rs1536482 was associated with keratoconus in a candidate gene study which assessed 44 SNPs in or

nearby *COL5A1* in two independent case-control cohorts and a familial cohort from the USA ($p = 6.5 \times 10^{-3}$).[299] The same SNP showed suggestive association with keratoconus ($p = 2.6 \times 10^{-7}$) in the study by Lu and colleagues,[87] which included both our cohort of keratoconus patients and the USA cohort included in the discovery phase of the USA study described previously.[299] In contrast, a small replication study did not show a significant association at rs1536482 in an independent Australian cohort of keratoconus patients ($n = 157$; $p = 0.4$).[88] When these three studies were considered together in a recent meta-analysis, taking into account the overlapping cohorts, the association at rs1536482 with keratoconus surpassed genome-wide significance (2.5×10^{-9}).[300] It is also worth noting that this SNP has not shown association with keratoconus in additional small cohorts (108-210 cases) from non-Caucasian populations such as such as a Han Chinese cohort ($p = 0.3$),[159] a cohort from Saudi Arabia ($p = 0.4$),[301] and a recent study in a Czech cohort ($p = 0.07$).[302] While these findings could be attributed to lack of power due to the small number of keratoconus patients included, it is also possible that the functional variant is absent or considerably rarer in these populations and therefore this locus may not play a large role in keratoconus susceptibility within these populations.

Our Australian keratoconus patients of European descent have been pivotal in the identification of the vast majority of previously known keratoconus-associated loci.[84, 86, 87] It was therefore unsurprising that two of the CCT-associated SNPs (rs2755238 and rs66720556), located at well-known keratoconus-associated loci, were significantly associated with keratoconus in Aim 1 of this study. This finding was however encouraging as both these SNPs had hard-typed genotyping data available for the cases and imputed genotypes for the controls, highlighting a key limitation for this stage of the study. Hard-typed genotype data were available for all 72 SNPs in the cases and approximately half of the SNPs in control cohort, with imputed genotype data available for remaining SNPs in the controls. Without consistency across the case and control data it is possible that any identified associations were due to the differences in the data, rather than true associations with disease. Further evidence that the significant associations were not affected by these differences include the fact that two SNPs at the same locus (*RXRA-COL5A1*), one with, and one without, imputed genotypes in the control data also surpassed the Bonferroni-corrected significance threshold. Together, these findings suggest that the imputation in the control cohort was of high quality and had minimal influence on this association study. This limitation was eliminated in Aim 2 of the study as genome-wide imputation was performed for both the cases and controls and used to fine-map keratoconus-associated loci. This second stage of the study also allowed for confirmation of the associations identified in Aim 1, albeit in a smaller cohort.

The fine-mapping analysis identified strong association signals in our case-control cohort at five of the fine-mapped keratoconus-associated loci: *FOXO1*, *FNDC3B*, *MPDZ*, *RXRA-COL5A1*, and the novel locus *KERA-LUM-DCN*. This experiment was used to identify trends in the data and explore the extent of the association signals at each keratoconus-associated locus, rather than focus on the significance of the associations in a small cohort with low power. The only keratoconus-loci that didn't display a clear

association peak was the *RAB3GAP1* locus. This locus was originally implicated in keratoconus in cohorts from the USA,[85] and reached genome-wide significance following replication and meta-analysis with our cohort.[86] While the top SNP identified in the present study, rs4954218, was the same SNP that was previously reported,[85] no surrounding SNPs were in high LD with this SNP. In fact, only five SNPs within the fine-mapped region obtained pairwise r^2 values with rs4954218 above 0.2, but all were below 0.4. Despite this, the association at rs4954218 observed in the present study ($p = 0.002$) was similar to our previously published findings ($p = 0.003$),[86] with the difference likely due to slight differences in the included cases and controls. In isolation, our data suggests that this keratoconus-associated locus is a false positive, however, our cohort only contributed marginally to the genome-wide association with keratoconus,[86] with the majority of the signal coming from the USA cohorts.[85] Based on these findings, this region was not pursued further in our cohort of keratoconus patients. Further studies in large independent cohorts should be conducted to confirm the role of rs4954218 in keratoconus susceptibility, and if the association is a true positive, investigation of this locus in the cohorts from the USA may help elucidate the functional variant and the potential mechanism of disease.

To aid the identification the functional variants at keratoconus-associated loci, keratoconus patients carrying the risk-associated allele at the top SNP at one or more of the selected loci were re-sequenced in Aim 3, along with a small cohort of unaffected controls. As the cases were specifically selected based on their genotype at risk-associated loci, the risk-associated haplotypes were artificially enriched in this group. For this reason, no formal statistics based on the frequency of variants were conducted, however, this artificial enrichment was exploited when prioritising variants in high LD with the top SNP. In contrast to the cases, the controls were not selected based on genotype and therefore variants identified within this cohort should resemble the normal population frequency. This allowed variants that were observed at similar frequencies in both the cases and controls to be excluded as they were not likely to be carried on the disease-associated haplotype or contribute to keratoconus risk. Variants were also excluded during filtering if they were not enriched, albeit artificially, in our cases compared to the maximum population frequency reported in gnomAD as there is no evidence that any of the populations in this database have a higher prevalence of keratoconus than individuals of European descent, and therefore, these variants were considered unlikely to contribute to disease.

As is the nature of targeted enrichment, some regions were not included in the design or obtained poor coverage (such as GC-rich regions and repetitive sequences) and therefore variants located in these regions were not able to be assessed in the re-sequencing study. While on average the remaining regions had sufficient depth for variant calling, not all individuals obtained high confidence genotypes at all sites due to variations in the read depth and genotype quality. To ensure all prioritised variants were real, all variant sites were required to obtain high confidence genotype calls in at least 50% of the re-sequenced cohort, including multiple high confidence variant calls. While this method inevitably included some low confidence variant calls and this was likely to affect the variant frequency estimates

in our cases and controls, these estimates were already limited by the small cohort size, particularly the control group, and the biased selection method used for the cases. We therefore suggest that the essential next step for any variants of interest identified in Aim 3 should be genotyped in a much larger cohort of cases and controls to determine more accurate variant frequencies in these groups and confirm the enrichment in cases compared to controls.

Due to the requirement for prioritised variants to be observed more than once in the re-sequencing data and limited size of the cohort, particularly the control group, the re-sequencing study was biased against rare variants (MAF < 1%). Despite this, the filtering strategy was designed to include low frequency variants (MAF < 5%) as they are currently difficult to impute accurately,[303] but are hypothesised to play an important role in the susceptibly to complex disease.[304, 305] To ensure both low frequency and common variants in high LD with the top SNP were prioritised, D' values were used instead of r^2 values which are heavily influenced by differences in allele frequencies. In addition, only variants with a D' value with a corresponding LOD score of at least 2 were included to ensure the LD estimates were of high confidence.

To further investigate the highly prioritised variants identified in the re-sequencing analysis (Aim 3), *in silico* tools such as CADD and FATHMM, as well as ENCODE data and eQTL data from the GETx database were used to identify putatively functional variants. All of the highly prioritised variants identified in this stage of the study were non-protein-coding variants and across all five loci, five highly prioritised variants obtained a CADD score above 10, and only two scored within the pathogenic range using the FATHMM-MKL algorithm (0.5 – 1.0). While the low scores obtained for these variants may suggest that they are unlikely to be functional, it also could reflect the limited understanding of the role non-coding regions of the genome within the broad field of genetics.

A compelling putatively functional variant, rs79728429, was identified as at the *FOXO1* locus. This variant was highly prioritised following re-sequencing and further investigations revealed that rs79728429 is located within a likely enhancer region and obtained a scaled CADD score indicative of pathogenicity. Furthermore, we propose the newly discovered gene, *AL133318.1*, as the target gene for rs79728429. This is based on the observation that rs7333246, 132 bp downstream of rs79728429 and located within the same enhancer region, is a known eQTL for *AL133318.1* in five tissues assessed in the GETx database, suggesting that this enhancer is important in the appropriate expression of *AL133318.1* in specific tissues. One of the tissues associated with *AL133318.1* expression at rs7333246 was sun exposed skin, which is the most relevant tissue for the cornea available in the GETx database, given that the database does not include any ocular tissues. From this it was hypothesised that the enhancer region is important for corneal expression of *AL133318.1*, and that the variant at rs79728429 alters this expression, contributing to an increased susceptibility to keratoconus. This gene has only recently been identified and therefore is not present in key databases such as the Ocular Tissue Database

(https://genome.uiowa.edu/otdb/), therefore determining corneal expression of *AL133318.1* should be a high priority. If AL133318.1 is expressed in corneal tissue, functional analyses to determine the normal function of the protein and characterise any differences in keratoconic corneas should be conducted, including determining cellular/extracellular localisation and identifying interactions (with other proteins, RNA molecules, or DNA). Combined with bioinformatic analyses such as determining homology to other proteins, identifying key structural features, domains and potential targets, this would greatly improve our understanding of this protein in both normal biology and any potential role in disease.

5.8 CONCLUSION

Across three aims, this study thoroughly investigated the role of common SNPs and small indels in keratoconus susceptibility in our cohort of Australian keratoconus patients of European descent. Using a strategic endophenotype approach, a novel keratoconus-association was identified at rs3759221 at the *KERA-LUM-DCN* locus (Aim 1). To investigate the extent and strength of the association signal, this novel locus, along with five well-known keratoconus loci (*RAB3GAP1*, *FNDC3B*, *MPDZ-NFIB*, *RXRA-COL5A1* and *FOXO1*) were fine-mapped in Aim 2. This analysis showed strong association signals for all loci except *RAB3GAP1*. The remaining five loci were selected for re-sequencing in Aim 3. This work identified putatively functional variants at each locus, and for the first time in keratoconus genetics, proposed rs79728429 as a functional variant at the previously identified *FOXO1* locus. As the variant was located within an enhancer region that also contained a known eQTL for *AL133318.1* we went on to hypothesise that rs79728429 alters the expression of *AL133318.1* and that this is the mechanism of keratoconus-susceptibility at this locus. While the variant and gene require focused functional analysis to confirm the potential role in keratoconus susceptibility and pathogenesis, these novel findings are an essential foundation for elucidating the mechanisms of disease and key disease processes. Overall, the work outlined in this chapter clearly demonstrates that non-protein-coding variation is important in keratoconus susceptibility.

CHAPTER 6: FINAL DISCUSSION AND CONCLUSIONS

6.1 SUMMARY OF THE FINDINGS

It has long been hypothesised that rare protein-coding variants play a significant role in keratoconus development and pathogenesis, particularly in families with strong Mendelian inheritance patterns of disease. In contrast, genome-wide association studies (GWAS) have identified common non-coding variants associated with keratoconus susceptibility. To address this dichotomy, this dissertation employed multiple methodologies to elucidate variants involved in keratoconus susceptibility under two distinct, yet complementary hypotheses.

The first hypothesis – that rare, highly penetrant protein-coding variants contribute to keratoconus development – was addressed using both a case-control study and a family-based study. Chapter 3 described the case-control study, which demonstrated that rare coding variants in 22 literature-based candidate genes for keratoconus were unlikely to play a substantial role in disease. The analysis of two families with multiple cases of keratoconus was described in Chapter 4 and identified two novel linkage regions and replicated a third. One of the families displayed likely digenic inheritance of keratoconus; with two linkage regions demonstrating equal evidence of association with disease and all affected individuals carrying both disease-associated haplotypes. While no rare protein-coding variants were considered putatively disease-causing in either family, non-protein-coding putatively disease-causing variants were identified and prioritised, including a compelling variant located in and untranslated region (UTR) of *SMOX* (c.-224C>T) in one of the families. From this, we propose *SMOX* as a novel keratoconus-candidate gene and both the specific variant, and the gene, warrant further genetic analysis and functional studies to confirm their role in disease. While it is possible that rare protein-coding variants located within both the candidate genes in Chapter 3 and the linkage regions in the families in Chapter 4 were missed due to biases in the sequencing technologies and platforms or the informatic tools used, these studies did not support the original hypothesis and instead suggest that alternative hypotheses should be explored to aid the identification of specific genetic factors involved in keratoconus susceptibility.

The second hypothesis states that variants associated with keratoconus indicate haplotypes that harbour functional variants which directly contribute to keratoconus susceptibility. Using case-control cohorts, Chapter 5 addressed this hypothesis in three sequential aims. This work led to the identification of a novel keratoconus-susceptibility locus, rs3759221, at *KERA-LUM-DCN* and identified and prioritised putatively-functional variants at this locus, as well as at four previously published keratoconus-associated loci. From this work, we propose rs79728429 as a novel functional variant at the rs2721051 (*FOXO1*) locus. We further hypothesise that *AL133318.1* is a regulatory target of rs79728429 as the SNP is located 133 bp downstream of an established expression quantitative trait loci (eQTL) for

AL133318.1 (rs7333246) and both variants are located within the same enhancer region. Based on these findings, rs79728429 and *AL133318.1* should be further investigated to confirm their role in keratoconus susceptibility. Overall, the findings from this chapter indicate that non-coding variation is likely to play an important role in keratoconus susceptibility.

6.2 A LIMITED ROLE FOR RARE PROTEIN-CODING VARIATION IN KERATOCONUS SUSCEPTIBILITY

In the family study described in Chapter 4, no rare protein-coding variants were classified as putatively disease-causing variants in either of the families. In fact, only one exonic variant was identified within the region of homozygosity in KSA197; however, this variant was too common to account for disease. Similarly, 23 protein-coding variants segregated within the two linkage regions identified in KCNSW01, but only one was rare. This single rare variant, however, wasn't investigated further as it did not affect the amino acid sequence of the protein and obtained low predictions of deleteriousness and pathogenicity. While these findings did not support the original hypothesis, they are consistent with previous publications. To date, only one linkage study has identified a rare putatively disease-causing variant within a protein-coding region, specifically in the dedicator of cytokinesis 9 gene (*DOCK9*; rs191047852),[29, 64] but further analyses are required to confirm the role of this variant in disease. In contrast, two published family studies have proposed non-protein-coding variants as the likely disease-causing, including a variant located within a non-protein-coding RNA gene, *mir184*, that co-segregates with disease in two unrelated families.[65, 66, 77, 143] Studies of families with keratoconus have largely restricted their search for likely-causative variants to those that are located within protein-coding regions of candidate genes that map to linkage regions, however, the vast majority of families in the literature remain unsolved.[28, 71-76, 78-80, 113] Together, these findings suggest rare protein-coding variants are unlikely to play a substantial role in keratoconus development, even in families with multiple cases and strong Mendelian inheritance patterns, and that alternative hypotheses should be investigated.

Further evidence for a limited role of protein-coding variants in keratoconus is presented in Chapters 3 and 5. Chapter 3 described a large study that screened 22 genes that had been proposed as candidates for keratoconus based on genes located within linkage-regions identified in family studies, the known function of the gene, or proximity to keratoconus-associated loci. This study demonstrated no difference in the frequency of rare protein-coding variants that were predicted to be potentially pathogenic between keratoconus patients compared to controls, suggesting that rare protein-coding variants within these genes do not play a substantial role in disease. Though it was not the primary focus of the analysis, the re-sequencing experiment presented in Chapter 5 further highlighted that protein-coding variants located within genes near GWAS hits do not to contribute to keratoconus susceptibility as none had the potential to be the functional variant across the five loci assessed. In fact, no protein-coding variants

were identified in cases or controls at three of the re-sequenced loci – *MPDZ-NFIB*, *RXRA-COL5A1* and *KERA-LUM-DCN* – despite the capture of multiple genes. Only three protein-coding variants were identified at the *FNDC3B* locus, and five within the *FOXO1* locus, however these variants were either too rare, or not in high linkage disequilibrium (LD; with high confidence) with the risk alleles and weren't prioritised. This phenomenon is not unique to keratoconus, with a recent study that investigated more than 6,000 significant associations from 920 GWAS for complex diseases and traits demonstrating that less than 5% of the associations were in coding regions and that while more than 41% of the trait-associated SNPs were located within introns, less than 11% of these variants were in high LD ($r^2 \geq 0.8$) with any coding variants.[306] Based on our findings, we now hypothesise that non-coding variants play a substantial role in keratoconus susceptibility by regulating or altering the expression of key genes. For this reason, we propose that future keratoconus studies should ensure appropriate capture and analysis of non-coding variation.

Another hypothesis worthy of future investigation in keratoconus cohorts is that structural and copy number variation contribute to keratoconus susceptibility. Structural and copy number variants include large insertions or deletions, as well as, more complex changes in the DNA sequence such as inversions or tandem duplications. Due to the large genomic regions affected by these variants, purpose-developed variant calling algorithms are required to call these variants from short read-sequencing data and therefore these types of variants may be present within protein-coding regions but have not yet been investigated. Evidence from obesity studies support the hypothesis that rare structural variants are important in complex disease,[307, 308] however it is worth noting that the associated regions were identified in case-control studies involving individuals with extreme phenotypes, with no overlapping signal identified in large GWAS.[307] For this reason, the investigation of such variants in keratoconus families should be of primary interest. This is further supported by the fact that copy number variations have been previously linked to other ocular diseases, including severe developmental ocular malformations[309] and a partial gene duplication was found to co-segregate with isolated congenital cataract in a large family.[310] Therefore, to extend the work presented in Chapter 4, these variants should be called from the WGS data generated for the families and analysed. To the best of our knowledge, structural and copy number variation has not previously been explored in keratoconus cohorts, and therefore, the data generated in this project presents a novel opportunity to explore the role of these variants in keratoconus susceptibility.

6.3 CHALLENGES WITH DETERMINING THE FUNCTIONAL IMPACT OF NON-PROTEIN-CODING VARIANTS

Non-protein-coding regions make up approximately 97% of the human genome and include the untranslated regions of genes; introns; non-coding RNA molecules; and intergenic regions that harbour

a complex suite of regulatory elements such as promotors, enhancers, silencers, insulators, and many other genomic features that we do not yet fully understand. Non-coding regions are also important for the appropriate the structure of DNA, allowing for looping of distal regulatory elements into close proximity to the target genes, as well as regulating the access of transcription factors and proteins to the DNA. The challenge with non-coding variation is identifying which parts of the DNA are likely to be important in disease, a task that is further complicated by the fact that these regions may only be functional in specific cell types or for transient periods of time. It is anticipated however, that our ability to prioritise and identify functional non-coding variants will continue to improve as the field develops a deeper understanding of the function of non-coding regions, and as public datasets continue to grow.

To aid the prioritisation of non-protein-coding variants in the present study, highly prioritised variants were manually assessed using tracks available on the University of California Santa Cruz (UCSC) Genome Browser, primarily by visualising data from the Encyclopedia of DNA Elements (ENCODE). ENCODE is a public research project that aims to identify functional elements within the human genome using a wide variety of assays and methods and in a selection of cell lines and tissue types. These data were used to identify likely regulatory elements such as enhancers and were complemented by data available from the Genotype-Tissue Expression Project (GTEx). GTEx aims to build a comprehensive resource containing eQTLs, identified in non-diseased tissues across almost 1000 individuals. Given the available data, non-coding variants located within likely enhancer regions were prioritised in both Chapters 4 and 5, though this method had a few caveats. It is important to highlight that neither of these databases include corneal tissue, and therefore these data may not be representative of the cornea. Sun exposed skin is likely the most relevant tissue in the GTEx database and data from epidermal keratinocytes (NHEK) is likely to be the most relevant in the ENCODE data. Another limitation of this strategy was that variants located in enhancer regions or DNaseI hypersensitivity regions that were observed across multiple cell types were more highly prioritised than those identified in fewer cell types as it provided more confidence in the prediction of the element. This methodology therefore biases top ranked variants toward those that affect non-tissue specific regulatory elements and therefore variants within these regions may be less likely to cause a tissue-specific phenotype such as keratoconus. Despite these limitations, by combining these data along with *in silico* predictions of pathogenicity and deleteriousness, this dissertation proposed two non-protein-coding variants as putatively functional variants in keratoconus susceptibility: a variant located in the 5' UTR of *SMOX* (c.-224C>T) was found to co-segregate with disease in a large family with multiple cases of keratoconus and a novel putatively-functional variant (rs79728429) was identified at a common keratoconus-susceptibility locus near *FOXO1*.

6.4 STRENGTHS AND LIMITATIONS OF CASE-CONTROL STUDIES

One of the major strengths of this dissertation was our large cohort of keratoconus patients consisting of over 620 cases. To reduce the influence of population stratification, ethnically matched controls were selected for the case-control studies. Another strength was the use of older control cohorts consisting of individuals who were unaffected by keratoconus, where possible. While keratoconus can develop at any age, it typically develops during adolescence or early adulthood, and therefore keratoconus cohorts are generally skewed toward younger participants. Therefore, the use of substantially older control groups reduces the risk that these individuals will later develop keratoconus and minimises the miss-classification of these individuals as controls. It is important to note, however, that this strength does not extend to the Anglo-Australasian Osteoporosis Consortium cohort as these individuals were not specifically examined for eye disease. Together, the large case cohort and strategically selected control cohorts facilitated the replication of previously published gene-screen studies in the largest cohort to date in Chapter 3 and thoroughly investigate keratoconus-associated loci in Chapter 5.

While the most appropriate control cohorts available were selected for analysis throughout this dissertation, some of the control individuals were affected by eye diseases other than keratoconus, which represents a potential limitation of the studies. The initial concern was around the potential of glaucoma as a confounding factor given both keratoconus and glaucoma share thin CCT as a risk factor.[31, 311-314] Most notably, this may have affected the analysis of *ZNF469* in Chapter 3 as the screened control cohort consisted of individuals from the Australian and New Zealand Registry of Advanced Glaucoma, who were almost all affected by primary open-angle glaucoma (POAG). In addition, the control cohorts from Anglo-Australasian Osteoporosis Consortium and the Blue Mountains Eye Study were expected to have the normal population frequencies of eye disease, including glaucoma. Despite the fact that CCT is phenotypic risk factor for glaucoma, a large study of 22 extended pedigrees enriched for POAG demonstrated that there was no significant genetic correlation between CCT and POAG-risk ($p = 0.27$).[315] This finding is further supported by two studies that assessed CCT-associated variants in cohorts of keratoconus and POAG patients, which led to the identification of novel keratoconus-associated loci (including the *ZNF469* locus), but neither study identified any variants associated with POAG.[87, 249] From these studies we propose that CCT and POAG, and by extension POAG and keratoconus, do not share substantial genetic overlap and therefore suggest that the inclusion of control individuals with glaucoma are unlikely to affect our findings.

While quite a few family-based studies have included families from other ethnicities, all published keratoconus GWAS signals have been identified in European populations, despite a higher prevalence of disease in Asian populations.[19] Differences in prevalence of disease across populations may be indicative of a region-specific environmental factors, however, could also be driven by genetic risk factors. As all the case-control studies presented in this dissertation were conducted in cohorts of

Australians of European descent there is a possibility that some of the findings from this work may be population-specific. Some of the findings, particularly those presented in Chapter 5, may not be applicable to keratoconus more broadly due to differences in variant frequencies and LD structures across populations. For example, in the recent cross-ancestry meta-analysis for CCT, two loci identified in Europeans were monomorphic in the Asian cohorts and therefore do not contribute to CCT in this population.[249] While population-specific susceptibility variants are likely to exist, the underlying biological pathways and mechanisms involved in keratoconus should converge across populations. Considering our limited understanding of biological underpinnings of keratoconus, studying keratoconus genetics in other populations would lead to a more complete understanding of specific genetic factors involved in the disease and may help to elucidate key biological pathways or mechanisms of disease. Better understanding of this would aid the development of biomarkers for early diagnosis and less invasive therapies, improve management strategies for patients and ultimately improve outcomes for keratoconus patients.

6.5 FUTURE DIRECTIONS

This study identified two novel keratoconus-candidate genes of great interest: *SMOX* and *AL133318.1*. The SMOX protein is induced by inflammation,[211-213] plays a role in apoptosis and the cellular response to both oxidative stress and ultraviolet radiation.[210] As keratoconus is characterised by progressive corneal thinning, likely due to the loss of stromal keratocytes and extracellular matrix degeneration,[316] these biological pathways are highly plausible in keratoconus development and pathogenesis. By extension, it has been hypothesised that variants that alter the expression of *SMOX* in the cornea are involved in keratoconus susceptibility and thus further genetic studies and functional analyses should be explored. The identification of the specific variant identified in KCNSW01 (c.-224C>T) in additional keratoconus patients would be strong evidence for the involvement of this variant and gene in the disease process, however, considering the rarity of the variant (maximum frequency of 0.4% in the African population of gnomAD) this is unlikely. For this reason, we propose a gene-based association study – similar to that which was conducted in Chapter 3 but ensuring that non-protein-coding regulatory regions are included – to determine if there is an enrichment of c.-224C>T, and other similar putatively disease-causing variants, in cases compared to population controls. In contrast, there is very little known about the single exon gene, *AL133318.1*, apart from the fact that it encodes a polypeptide consisting of 62 amino acids. For this reason, determining if *AL133318.1* is expressed in the cornea should be of primary focus. Any differences in expression between healthy and keratoconic corneas should also be determined. If *AL133318.1* is expressed in corneal tissue, laboratory experiments to determine the localisation of the protein and any interactions, including protein-protein interactions, to develop an understanding of the function of the protein. Considering the likely digenic inheritance in KCNSW01,

the family in which the *SMOX* variant was identified, similar analyses would also be worth pursuing for *SMOX*, as these analyses in corneal tissue may help elucidate candidate genes and aid the prioritisation of putatively disease-causing variants located within the 17q12 linkage region.

As both the *SMOX* 5' UTR variant (c.-224C>T) and rs79728429 near *AL133318.1* are hypothesised to have a regulatory role, the expression of SMOX and AL133318.1 should be explored in corneal tissue or relevant cell lines with and without the relevant variants and the results compared. This could be achieved using corneal tissue or through an induced pluripotent stem cell model, from variant carriers and individuals without the variants of interest. This type of analysis may be more feasible for the more common variant, rs79728429, as the rarity of the c.-224C>T in the 5' UTR of *SMOX* realistically limits this type of analysis to the family members themselves. Despite this, family members of KCNSW01 have already demonstrated a willingness to participate in research and interest in understanding their disease, and therefore, may be willing to provide tissue samples. This would allow for the analysis of the *SMOX* variant, and other variants located within the two linkage regions in KCNSW01, alongside the individuals' full genetic background. Alternatively, gene editing technologies such as clustered regularly interspaced short palindromic repeats (CRISPR)/cas9 may be used to insert the variants of interest into cells that do not natively carry the variant, allowing for the investigation of the impact of the variants of interest on gene expression without the need for participants' samples.

6.6 FINAL CONCLUSIONS

This dissertation used both case-control and family-based study designs to elucidate variants involved in keratoconus susceptibility in a cohort of Australians of European ancestry. This work identified two novel linkage regions, replicated a third and proposed putatively disease-causing variants for further investigation in two families. Further analysis of the keratoconus-associated regions led to the identification of a novel keratoconus-candidate gene, *SMOX*, which warrants further investigation to confirm its role in this disease. In the case-control studies a novel keratoconus-susceptibility locus at *KERA-LUM-DCN* was identified and we proposed a functional variant (rs79728429) and novel target gene (*AL133318.1*) at a previously identified keratoconus-associated locus: rs2721051 (*FOXO1*). This dissertation also demonstrated that rare coding variants in 22 candidate genes for keratoconus were unlikely to play a substantial role in disease. Overall, this work highlights the limited contribution of rare protein-coding variation in keratoconus and suggests that non-coding or regulatory variants are likely to play a substantial role in disease susceptibility. This work also highlights the truly complex nature of keratoconus genetics, even in families with apparently Mendelian inheritance patterns of disease.

REFERENCES

1. Lucas SEM, Zhou T, Blackburn NB, *et al.* Rare, potentially pathogenic variants in ZNF469 are not enriched in keratoconus in a large Australian cohort of European descent. Investigative ophthalmology & visual science. 2017;58(14):6248-56.
2. Lucas SEM, Zhou T, Blackburn NB, *et al.* Rare, potentially pathogenic variants in 21 keratoconus candidate genes are not enriched in cases in a large Australian cohort of European descent. PloS one. 2018;13(6):e0199178.
3. Jonas JB, Nangia V, Matin A, *et al.* Prevalence and associations of keratoconus in rural maharashtra in central India: the central India eye and medical study. American journal of ophthalmology. 2009;148(5):760-5.
4. Kennedy RH, Bourne WM, Dyer JA. A 48-year clinical and epidemiologic study of keratoconus. American journal of ophthalmology. 1986;101(3):267-73.
5. Nielsen K, Hjortdal J, Aagaard Nohr E, *et al.* Incidence and prevalence of keratoconus in Denmark. Acta ophthalmologica Scandinavica. 2007;85(8):890-2.
6. Pearson AR, Soneji B, Sarvananthan N, *et al.* Does ethnic origin influence the incidence or severity of keratoconus? Eye (London, England). 2000;14 (Pt 4):625-8.
7. Tanabe U, Fujiki K, Ogawa A, *et al.* [Prevalence of keratoconus patients in Japan]. Nippon Ganka Gakkai zasshi. 1985;89(3):407-11.
8. Assiri AA, Yousuf BI, Quantock AJ, *et al.* Incidence and severity of keratoconus in Asir province, Saudi Arabia. The British journal of ophthalmology. 2005;89(11):1403-6.
9. Millodot M, Shneor E, Albou S, *et al.* Prevalence and associated factors of keratoconus in Jerusalem: a cross-sectional study. Ophthalmic epidemiology. 2011;18(2):91-7.
10. Waked N, Fayad AM, Fadlallah A, *et al.* [Keratoconus screening in a Lebanese students' population]. Journal francais d'ophtalmologie. 2012;35(1):23-9.
11. Xu L, Wang YX, Guo Y, *et al.* Prevalence and associations of steep cornea/keratoconus in Greater Beijing. The Beijing Eye Study. PloS one. 2012;7(7):e39313.
12. Hashemi H, Beiranvand A, Khabazkhoob M, *et al.* Prevalence of keratoconus in a population-based study in Shahroud. Cornea. 2013;32(11):1441-5.
13. Hashemi H, Khabazkhoob M, Fotouhi A. Topographic Keratoconus is not Rare in an Iranian population: the Tehran Eye Study. Ophthalmic epidemiology. 2013;20(6):385-91.
14. Hashemi H, Khabazkhoob M, Yazdani N, *et al.* The prevalence of keratoconus in a young population in Mashhad, Iran. Ophthalmic & physiological optics : the journal of the British College of Ophthalmic Opticians (Optometrists). 2014;34(5):519-27.
15. Cozma I, Atherley C, James NJ. Influence of ethnic origin on the incidence of keratoconus and associated atopic disease in Asian and white patients. Eye (London, England). 2005;19(8):924-5; author reply 5-6.
16. Georgiou T, Funnell CL, Cassels-Brown A, *et al.* Influence of ethnic origin on the incidence of keratoconus and associated atopic disease in Asians and white patients. Eye (London, England). 2004;18(4):379-83.
17. Ota R, Fujiki K, Nakayasu K. [Estimation of patient visit rate and incidence of keratoconus in the 23 wards of Tokyo]. Nippon Ganka Gakkai zasshi. 2002;106(6):365-72.
18. Ziaei H, Jafarinasab MR, Javadi MA, *et al.* Epidemiology of keratoconus in an Iranian population. Cornea. 2012;31(9):1044-7.
19. Kok YO, Tan GF, Loon SC. Review: keratoconus in Asia. Cornea. 2012;31(5):581-93.
20. Owens H, Gamble GD, Bjornholdt MC, *et al.* Topographic indications of emerging keratoconus in teenage New Zealanders. Cornea. 2007;26(3):312-8.
21. Zadnik K, Steger-May K, Fink BA, *et al.* Between-eye asymmetry in keratoconus. Cornea. 2002;21(7):671-9.
22. Saad A, Gatinel D. Topographic and tomographic properties of forme fruste keratoconus corneas. Investigative ophthalmology & visual science. 2010;51(11):5546-55.
23. Ye C, Ng PK, Jhanji V. Optical quality assessment in normal and forme fruste keratoconus eyes with a double-pass system: a comparison and variability study. The British journal of ophthalmology. 2014.

24. Leoni-Mesplie S, Mortemousque B, Touboul D, *et al.* Scalability and severity of keratoconus in children. American journal of ophthalmology. 2012;154(1):56-62.e1.
25. Mastropasqua L. Collagen cross-linking: when and how? A review of the state of the art of the technique and new perspectives. Eye and vision (London, England). 2015;2:19.
26. Tuft SJ, Moodaley LC, Gregory WM, *et al.* Prognostic factors for the progression of keratoconus. Ophthalmology. 1994;101(3):439-47.
27. Bareja U, Vajpayee RB. Posterior keratoconus due to iron nail injury--a case report. Indian journal of ophthalmology. 1991;39(1):30.
28. Brancati F, Valente EM, Sarkozy A, *et al.* A locus for autosomal dominant keratoconus maps to human chromosome 3p14-q13. Journal of medical genetics. 2004;41(3):188-92.
29. Gajecka M, Radhakrishna U, Winters D, *et al.* Localization of a gene for keratoconus to a 5.6-Mb interval on 13q32. Investigative ophthalmology & visual science. 2009;50(4):1531-9.
30. Gordon-Shaag A, Millodot M, Shneor E, *et al.* The genetic and environmental factors for keratoconus. BioMed research international. 2015;2015:795738.
31. Doughty MJ, Zaman ML. Human corneal thickness and its impact on intraocular pressure measures: a review and meta-analysis approach. Survey of ophthalmology. 2000;44(5):367-408.
32. Hassell JR, Birk DE. The molecular basis of corneal transparency. Experimental eye research. 2010;91(3):326-35.
33. Quantock AJ, Young RD. Development of the corneal stroma, and the collagen-proteoglycan associations that help define its structure and function. Developmental dynamics : an official publication of the American Association of Anatomists. 2008;237(10):2607-21.
34. Nakayasu K, Tanaka M, Konomi H, *et al.* Distribution of types I, II, III, IV and V collagen in normal and keratoconus corneas. Ophthalmic research. 1986;18(1):1-10.
35. Morishige N, Takagi Y, Chikama T, *et al.* Three-dimensional analysis of collagen lamellae in the anterior stroma of the human cornea visualized by second harmonic generation imaging microscopy. Investigative ophthalmology & visual science. 2011;52(2):911-5.
36. Maurice DM. The structure and transparency of the cornea. The Journal of physiology. 1957;136(2):263-86.
37. Ernst BJ, Hsu HY. Keratoconus association with axial myopia: a prospective biometric study. Eye & contact lens. 2011;37(1):2-5.
38. Klyce SD. Chasing the suspect: keratoconus. The British journal of ophthalmology. 2009;93(7):845-7.
39. Romero-Jimenez M, Santodomingo-Rubido J, Wolffsohn JS. Keratoconus: a review. Contact lens & anterior eye : the journal of the British Contact Lens Association. 2010;33(4):157-66; quiz 205.
40. Rabinowitz YS. Keratoconus. Survey of ophthalmology. 1998;42(4):297-319.
41. Maharana PK, Sharma N, Vajpayee RB. Acute corneal hydrops in keratoconus. Indian journal of ophthalmology. 2013;61(8):461-4.
42. Klyce SD. Computer-assisted corneal topography. High-resolution graphic presentation and analysis of keratoscopy. Investigative ophthalmology & visual science. 1984;25(12):1426-35.
43. Maeda N, Klyce SD, Smolek MK, *et al.* Automated keratoconus screening with corneal topography analysis. Investigative ophthalmology & visual science. 1994;35(6):2749-57.
44. Rabinowitz YS, Rasheed K. KISA% index: a quantitative videokeratography algorithm embodying minimal topographic criteria for diagnosing keratoconus. Journal of cataract and refractive surgery. 1999;25(10):1327-35.
45. Randleman JB, Russell B, Ward MA, *et al.* Risk factors and prognosis for corneal ectasia after LASIK. Ophthalmology. 2003;110(2):267-75.
46. Randleman JB, Woodward M, Lynn MJ, *et al.* Risk assessment for ectasia after corneal refractive surgery. Ophthalmology. 2008;115(1):37-50.
47. Barnett M, Mannis MJ. Contact lenses in the management of keratoconus. Cornea. 2011;30(12):1510-6.
48. Rathi VM, Mandathara PS, Dumpati S. Contact lens in keratoconus. Indian journal of ophthalmology. 2013;61(8):410-5.
49. Kelly TL, Coster DJ, Williams KA. Repeat penetrating corneal transplantation in patients with keratoconus. Ophthalmology. 2011;118(8):1538-42.

50. Kelly TL, Williams KA, Coster DJ. Corneal transplantation for keratoconus: a registry study. Archives of ophthalmology. 2011;129(6):691-7.
51. Williams KA, Muehlberg SM, Lewis RF, *et al.* How successful is corneal transplantation? A report from the Australian Corneal Graft Register. Eye (London, England). 1995;9 (Pt 2):219-27.
52. Williams KA LM, Jones VJ, Loh RS, Coster DJ The Australian Corneal Graft Registry 2012 Report. Adelaide, Australia; 2012.
53. Cassidy D, Beltz J, Jhanji V, *et al.* Recent advances in corneal transplantation for keratoconus. Clinical & experimental optometry : journal of the Australian Optometrical Association. 2013;96(2):165-72.
54. Bersudsky V, Blum-Hareuveni T, Rehany U, *et al.* The profile of repeated corneal transplantation. Ophthalmology. 2001;108(3):461-9.
55. Weisbrod DJ, Sit M, Naor J, *et al.* Outcomes of repeat penetrating keratoplasty and risk factors for graft failure. Cornea. 2003;22(5):429-34.
56. Yildiz EH, Hoskins E, Fram N, *et al.* Third or greater penetrating keratoplasties: indications, survival, and visual outcomes. Cornea. 2010;29(3):254-9.
57. Wollensak G, Spoerl E, Seiler T. Riboflavin/ultraviolet-a-induced collagen crosslinking for the treatment of keratoconus. American journal of ophthalmology. 2003;135(5):620-7.
58. Farjadnia M, Naderan M. Corneal cross-linking treatment of keratoconus. Oman journal of ophthalmology. 2015;8(2):86-91.
59. Ihalainen A. Clinical and epidemiological features of keratoconus genetic and external factors in the pathogenesis of the disease. Acta ophthalmologica Supplement. 1986;178:1-64.
60. Shneor E, Millodot M, Blumberg S, *et al.* Characteristics of 244 patients with keratoconus seen in an optometric contact lens practice. Clinical & experimental optometry : journal of the Australian Optometrical Association. 2013;96(2):219-24.
61. Weed KH, MacEwen CJ, Giles T, *et al.* The Dundee University Scottish Keratoconus study: demographics, corneal signs, associated diseases, and eye rubbing. Eye (London, England). 2008;22(4):534-41.
62. Wang Y, Rabinowitz YS, Rotter JI, *et al.* Genetic epidemiological study of keratoconus: evidence for major gene determination. American journal of medical genetics. 2000;93(5):403-9.
63. Kriszt A, Losonczy G, Berta A, *et al.* Segregation analysis suggests that keratoconus is a complex non-mendelian disease. Acta ophthalmologica. 2014.
64. Czugala M, Karolak JA, Nowak DM, *et al.* Novel mutation and three other sequence variants segregating with phenotype at keratoconus 13q32 susceptibility locus. European journal of human genetics : EJHG. 2012;20(4):389-97.
65. Hughes AE, Bradley DT, Campbell M, *et al.* Mutation altering the miR-184 seed region causes familial keratoconus with cataract. American journal of human genetics. 2011;89(5):628-33.
66. Nowak DM, Karolak JA, Kubiak J, *et al.* Substitution at IL1RN and deletion at SLC4A11 segregating with phenotype in familial keratoconus. Investigative ophthalmology & visual science. 2013;54(3):2207-15.
67. Karolak JA, Polakowski P, Szaflik J, *et al.* Molecular Screening of Keratoconus Susceptibility Sequence Variants in VSX1, TGFBI, DOCK9, STK24, and IPO5 Genes in Polish Patients and Novel TGFBI Variant Identification. Ophthalmic genetics. 2015:1-7.
68. Palamar M, Onay H, Ozdemir TR, *et al.* Relationship between IL1beta-511C>T and ILRN VNTR polymorphisms and keratoconus. Cornea. 2014;33(2):145-7.
69. Bykhovskaya Y, Caiado Canedo AL, Wright KW, *et al.* C.57 C > T Mutation in MIR 184 is Responsible for Congenital Cataracts and Corneal Abnormalities in a Five-generation Family from Galicia, Spain. Ophthalmic genetics. 2015;36(3):244-7.
70. Lechner J, Bae HA, Guduric-Fuchs J, *et al.* Mutational analysis of MIR184 in sporadic keratoconus and myopia. Investigative ophthalmology & visual science. 2013;54(8):5266-72.
71. Burdon KP, Coster DJ, Charlesworth JC, *et al.* Apparent autosomal dominant keratoconus in a large Australian pedigree accounted for by digenic inheritance of two novel loci. Human genetics. 2008;124(4):379-86.
72. Hutchings H, Ginisty H, Le Gallo M, *et al.* Identification of a new locus for isolated familial keratoconus at 2p24. Journal of medical genetics. 2005;42(1):88-94.

73. Bykhovskaya Y, Li X, Taylor KD, *et al.* Linkage Analysis of High-density SNPs Confirms Keratoconus Locus at 5q Chromosomal Region. Ophthalmic genetics. 2014.
74. Tang YG, Rabinowitz YS, Taylor KD, *et al.* Genomewide linkage scan in a multigeneration Caucasian pedigree identifies a novel locus for keratoconus on chromosome 5q14.3-q21.1. Genetics in medicine : official journal of the American College of Medical Genetics. 2005;7(6):397-405.
75. Li X, Rabinowitz YS, Tang YG, *et al.* Two-stage genome-wide linkage scan in keratoconus sib pair families. Investigative ophthalmology & visual science. 2006;47(9):3791-5.
76. Liskova P, Hysi PG, Waseem N, *et al.* Evidence for keratoconus susceptibility locus on chromosome 14: a genome-wide linkage screen using single-nucleotide polymorphism markers. Archives of ophthalmology. 2010;128(9):1191-5.
77. Hughes AE, Dash DP, Jackson AJ, *et al.* Familial keratoconus with cataract: linkage to the long arm of chromosome 15 and exclusion of candidate genes. Investigative ophthalmology & visual science. 2003;44(12):5063-6.
78. Tyynismaa H, Sistonen P, Tuupanen S, *et al.* A locus for autosomal dominant keratoconus: linkage to 16q22.3-q23.1 in Finnish families. Investigative ophthalmology & visual science. 2002;43(10):3160-4.
79. Hameed A, Khaliq S, Ismail M, *et al.* A novel locus for Leber congenital amaurosis (LCA4) with anterior keratoconus mapping to chromosome 17p13. Investigative ophthalmology & visual science. 2000;41(3):629-33.
80. Fullerton J, Paprocki P, Foote S, *et al.* Identity-by-descent approach to gene localisation in eight individuals affected by keratoconus from north-west Tasmania, Australia. Human genetics. 2002;110(5):462-70.
81. Lucas S. The aetiology of keratoconus from a genetic and epigenetic perspective. Hobart, Tasmania, Australia: University of Tasmania; 2014.
82. Daly MJ, Rioux JD, Schaffner SF, *et al.* High-resolution haplotype structure in the human genome. Nature genetics. 2001;29(2):229-32.
83. Shea J, Agarwala V, Philippakis AA, *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. Nature genetics. 2011;43(8):801-5.
84. Burdon KP, Macgregor S, Bykhovskaya Y, *et al.* Association of polymorphisms in the hepatocyte growth factor gene promoter with keratoconus. Investigative ophthalmology & visual science. 2011;52(11):8514-9.
85. Li X, Bykhovskaya Y, Haritunians T, *et al.* A genome-wide association study identifies a potential novel gene locus for keratoconus, one of the commonest causes for corneal transplantation in developed countries. Human molecular genetics. 2012;21(2):421-9.
86. Bae HA, Mills RA, Lindsay RG, *et al.* Replication and meta-analysis of candidate loci identified variation at RAB3GAP1 associated with keratoconus. Investigative ophthalmology & visual science. 2013;54(7):5132-5.
87. Lu Y, Vitart V, Burdon KP, *et al.* Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. Nature genetics. 2013;45(2):155-63.
88. Sahebjada S, Schache M, Richardson AJ, *et al.* Evaluating the association between keratoconus and the corneal thickness genes in an independent Australian population. Investigative ophthalmology & visual science. 2013;54(13):8224-8.
89. Heon E, Greenberg A, Kopp KK, *et al.* VSX1: a gene for posterior polymorphous dystrophy and keratoconus. Human molecular genetics. 2002;11(9):1029-36.
90. Bisceglia L, Ciaschetti M, De Bonis P, *et al.* VSX1 mutational analysis in a series of Italian patients affected by keratoconus: detection of a novel mutation. Investigative ophthalmology & visual science. 2005;46(1):39-45.
91. Eran P, Almogit A, David Z, *et al.* The D144E substitution in the VSX1 gene: a non-pathogenic variant or a disease causing mutation? Ophthalmic genetics. 2008;29(2):53-9.
92. Mok JW, Baek SJ, Joo CK. VSX1 gene variants are associated with keratoconus in unrelated Korean patients. Journal of human genetics. 2008;53(9):842-9.
93. Paliwal P, Singh A, Tandon R, *et al.* A novel VSX1 mutation identified in an individual with keratoconus in India. Molecular vision. 2009;15:2475-9.

94. Dash DP, George S, O'Prey D, *et al.* Mutational screening of VSX1 in keratoconus patients from the European population. Eye (London, England). 2010;24(6):1085-92.
95. Saee-Rad S, Hashemi H, Miraftab M, *et al.* Mutation analysis of VSX1 and SOD1 in Iranian patients with keratoconus. Molecular vision. 2011;17:3128-36.
96. De Bonis P, Laborante A, Pizzicoli C, *et al.* Mutational screening of VSX1, SPARC, SOD1, LOX, and TIMP3 in keratoconus. Molecular vision. 2011;17:2482-94.
97. Wang Y, Jin T, Zhang X, *et al.* Common single nucleotide polymorphisms and keratoconus in the Han Chinese population. Ophthalmic genetics. 2013;34(3):160-6.
98. Shetty R, Nuijts RM, Nanaiah SG, *et al.* Two novel missense substitutions in the VSX1 gene: clinical and genetic analysis of families with Keratoconus from India. BMC medical genetics. 2015;16:33.
99. Aldave AJ, Yellore VS, Salem AK, *et al.* No VSX1 gene mutations associated with keratoconus. Investigative ophthalmology & visual science. 2006;47(7):2820-2.
100. Liskova P, Ebenezer ND, Hysi PG, *et al.* Molecular analysis of the VSX1 gene in familial keratoconus. Molecular vision. 2007;13:1887-91.
101. Tang YG, Picornell Y, Su X, *et al.* Three VSX1 gene mutations, L159M, R166W, and H244R, are not associated with keratoconus. Cornea. 2008;27(2):189-92.
102. Stabuc-Silih M, Strazisar M, Hawlina M, *et al.* Absence of pathogenic mutations in VSX1 and SOD1 genes in patients with keratoconus. Cornea. 2010;29(2):172-6.
103. Udar N, Atilano SR, Brown DJ, *et al.* SOD1: a candidate gene for keratoconus. Investigative ophthalmology & visual science. 2006;47(8):3345-51.
104. Udar N, Atilano SR, Small K, *et al.* SOD1 haplotypes in familial keratoconus. Cornea. 2009;28(8):902-7.
105. Moschos MM, Kokolakis N, Gazouli M, *et al.* Polymorphism Analysis of VSX1 and SOD1 Genes in Greek Patients with Keratoconus. Ophthalmic genetics. 2013.
106. Al-Muammar AM, Kalantan H, Azad TA, *et al.* Analysis of the SOD1 Gene in Keratoconus Patients from Saudi Arabia. Ophthalmic genetics. 2015;36(4):373-5.
107. Bykhovskaya Y, Li X, Epifantseva I, *et al.* Variation in the lysyl oxidase (LOX) gene is associated with keratoconus in family-based and case-control studies. Investigative ophthalmology & visual science. 2012;53(7):4152-7.
108. Zhang J, Zhang L, Hong J, *et al.* Association of Common Variants in LOX with Keratoconus: A Meta-Analysis. PloS one. 2015;10(12):e0145815.
109. Dudakova L, Palos M, Jirsova K, *et al.* Validation of rs2956540:G>C and rs3735520:G>A association with keratoconus in a population of European descent. European journal of human genetics : EJHG. 2015;23(11):1581-3.
110. Hasanian-Langroudi F, Saravani R, Validad MH, *et al.* Association of Lysyl oxidase (LOX) polymorphisms with the risk of Keratoconus in an Iranian population. Ophthalmic genetics. 2014.
111. Piret SE, Gorvin CM, Pagnamenta AT, *et al.* Identification of a G-Protein Subunit-alpha11 Gain-of-Function Mutation, Val340Met, in a Family with Autosomal Dominant Hypocalcemia Type 2 (ADH2). Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research. 2016.
112. Guan T, Liu C, Ma Z, *et al.* The point mutation and polymorphism in keratoconus candidate gene TGFBI in Chinese population. Gene. 2012;503(1):137-9.
113. Li X, Bykhovskaya Y, Tang YG, *et al.* An association between the calpastatin (CAST) gene and keratoconus. Cornea. 2013;32(5):696-701.
114. Stabuc-Silih M, Ravnik-Glavac M, Glavac D, *et al.* Polymorphisms in COL4A3 and COL4A4 genes associated with keratoconus. Molecular vision. 2009;15:2848-60.
115. Wojcik KA, Synowiec E, Jimenez-Garcia MP, *et al.* Polymorphism of the transferrin gene in eye diseases: keratoconus and Fuchs endothelial corneal dystrophy. BioMed research international. 2013;2013:247438.
116. Synowiec E, Wojcik KA, Izdebska J, *et al.* Polymorphisms of the homologous recombination gene RAD51 in keratoconus and Fuchs endothelial corneal dystrophy. Disease markers. 2013;35(5):353-62.

117. Kim SH, Mok JW, Kim HS, *et al.* Association of -31T>C and -511 C>T polymorphisms in the interleukin 1 beta (IL1B) promoter in Korean keratoconus patients. Molecular vision. 2008;14:2109-16.
118. Lechner J. Genetic investigation of keratoconus. Belfast, Ireland: Queen's University Belfast; 2013.
119. Droitcourt C, Touboul D, Ged C, *et al.* A prospective study of filaggrin null mutations in keratoconus patients with or without atopic disorders. Dermatology (Basel, Switzerland). 2011;222(4):336-41.
120. Karolak JA, Rydzanicz M, Ginter-Matuszewska B, *et al.* Variant c.2262A>C in DOCK9 Leads to Exon Skipping in Keratoconus Family. Investigative ophthalmology & visual science. 2015;56(13):7687-90.
121. Lechner J, Dash DP, Muszynska D, *et al.* Mutational spectrum of the ZEB1 gene in corneal dystrophies supports a genotype-phenotype correlation. Investigative ophthalmology & visual science. 2013;54(5):3215-23.
122. Makinen VP, Parkkonen M, Wessman M, *et al.* High-throughput pedigree drawing. European journal of human genetics : EJHG. 2005;13(8):987-9.
123. Duncan EL, Danoy P, Kemp JP, *et al.* Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. PLoS genetics. 2011;7(4):e1001372.
124. Mitchell P, Smith W, Attebo K, *et al.* Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. Ophthalmology. 1996;103(10):1661-9.
125. McKenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
126. R Core Team. R: A Language and Environment for Statistical Computing. Viena, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.R-project.org/.
127. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164.
128. Sayers EW, Barrett T, Benson DA, *et al.* Database resources of the National Center for Biotechnology Information. Nucleic acids research. 2011;39(Database issue):D38-51.
129. Lek M, Karczewski KJ, Minikel EV, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.
130. Glusman G, Caballero J, Mauldin DE, *et al.* Kaviar: an accessible system for testing SNV novelty. Bioinformatics (Oxford, England). 2011;27(22):3216-7.
131. The Genomes Project C, Auton A, Abecasis GR, *et al.* A global reference for human genetic variation. Nature. 2015;526:68.
132. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research. 2001;11(5):863-74.
133. Adzhubei IA, Schmidt S, Peshkin L, *et al.* A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.
134. Kircher M, Witten DM, Jain P, *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014;46(3):310-5.
135. Shihab HA, Gough J, Cooper DN, *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human mutation. 2013;34(1):57-65.
136. Shihab HA, Rogers MF, Gough J, *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics (Oxford, England). 2015;31(10):1536-43.
137. Ferlaino M, Rogers MF, Shihab HA, *et al.* An integrative approach to predicting the functional effects of small indels in non-coding regions of the human genome. BMC bioinformatics. 2017;18(1):442.
138. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
139. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013;45(6):580-5.

140. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017;550:204.
141. Bisceglia L, De Bonis P, Pizzicoli C, *et al.* Linkage analysis in keratoconus: replication of locus 5q21.2 and identification of other suggestive Loci. Investigative ophthalmology & visual science. 2009;50(3):1081-6.
142. Rosenfeld JA, Drautz JM, Clericuzio CL, *et al.* Deletions and duplications of developmental pathway genes in 5q31 contribute to abnormal phenotypes. American journal of medical genetics Part A. 2011;155a(8):1906-16.
143. Dash DP, Silvestri G, Hughes AE. Fine mapping of the keratoconus with cataract locus on chromosome 15q and candidate gene analysis. Molecular vision. 2006;12:499-505.
144. Sahebjada S, Schache M, Richardson AJ, *et al.* Association of the hepatocyte growth factor gene with keratoconus in an Australian population. PloS one. 2014;9(1):e84067.
145. Lechner J, Porter LF, Rice A, *et al.* Enrichment of pathogenic alleles in the brittle cornea gene, ZNF469, in keratoconus. Human molecular genetics. 2014.
146. Vincent AL, Jordan CA, Cadzow MJ, *et al.* Mutations in the zinc finger protein gene, ZNF469, contribute to the pathogenesis of keratoconus. Investigative ophthalmology & visual science. 2014.
147. Davidson AE, Borasio E, Liskova P, *et al.* Brittle Cornea Syndrome ZNF469 mutation carrier phenotype and segregation analysis of rare ZNF469 variants in familial Keratoconus. Investigative ophthalmology & visual science. 2015;56(1):578-86.
148. Karolak JA, Gambin T, Rydzanicz M, *et al.* Evidence against ZNF469 being causative for keratoconus in Polish patients. Acta ophthalmologica. 2016.
149. Paliwal P, Tandon R, Dube D, *et al.* Familial segregation of a VSX1 mutation adds a new dimension to its role in the causation of keratoconus. Molecular vision. 2011;17:481-5.
150. Vincent AL, Jordan C, Sheck L, *et al.* Screening the visual system homeobox 1 gene in keratoconus and posterior polymorphous dystrophy cohorts identifies a novel variant. Molecular vision. 2013;19:852-60.
151. Bardak H, Gunay M, Yildiz E, *et al.* Novel visual system homeobox 1 gene mutations in Turkish patients with keratoconus. Genetics and molecular research : GMR. 2016;15(4).
152. Karolak JA, Polakowski P, Szaflik J, *et al.* Molecular Screening of Keratoconus Susceptibility Sequence Variants in VSX1, TGFBI, DOCK9, STK24, and IPO5 Genes in Polish Patients and Novel TGFBI Variant Identification. Ophthalmic genetics. 2016;37(1):37-43.
153. Tanwar M, Kumar M, Nayak B, *et al.* VSX1 gene analysis in keratoconus. Molecular vision. 2010;16:2395-401.
154. Abu-Amero KK, Hellani AM, Al Mansouri SM, *et al.* High-resolution analysis of DNA copy number alterations in patients with isolated sporadic keratoconus. Molecular vision. 2011;17:822-6.
155. Abu-Amero KK, Kalantan H, Al-Muammar AM. Analysis of the VSX1 gene in keratoconus patients from Saudi Arabia. Molecular vision. 2011;17:667-72.
156. Jeoung JW, Kim MK, Park SS, *et al.* VSX1 gene and keratoconus: genetic analysis in Korean patients. Cornea. 2012;31(7):746-50.
157. Dehkordi FA, Rashki A, Bagheri N, *et al.* Study of VSX1 mutations in patients with keratoconus in southwest Iran using PCR-single-strand conformation polymorphism/heteroduplex analysis and sequencing method. Acta cytologica. 2013;57(6):646-51.
158. Verma A, Das M, Srinivasan M, *et al.* Investigation of VSX1 sequence variants in South Indian patients with sporadic cases of keratoconus. BMC research notes. 2013;6:103.
159. Hao XD, Chen P, Chen ZL, *et al.* Evaluating the Association between Keratoconus and Reported Genetic Loci in a Han Chinese Population. Ophthalmic genetics. 2015;36(2):132-6.
160. Ng JB, Poh RY, Lee KR, *et al.* Visual System Homeobox 1 (VSX1) Gene Analysis in Keratoconus: Design of Specific Primers and DNA Amplification Protocols for Accurate Molecular Characterization. Clinical laboratory. 2016;62(9):1731-7.
161. Nejabat M, Naghash P, Dastsooz H, *et al.* VSX1 and SOD1 Mutation Screening in Patients with Keratoconus in the South of Iran. Journal of ophthalmic & vision research. 2017;12(2):135-40.

162. Kelly BJ, Fitch JR, Hu Y, *et al.* Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome biology. 2015;16:6.
163. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013;3(13033997).
164. Li H, Handsaker B, Wysoker A, *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-9.
165. DePristo MA, Banks E, Poplin R, *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011;43(5):491-8.
166. Van der Auwera GA, Carneiro MO, Hartl C, *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics. 2013;43:11.0.1-33.
167. Untergasser A, Nijveen H, Rao X, *et al.* Primer3Plus, an enhanced web interface to Primer3. Nucleic acids research. 2007;35(Web Server issue):W71-4.
168. Ye J, Coulouris G, Zaretskaya I, *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics. 2012;13:134.
169. Rohrbach M, Spencer HL, Porter LF, *et al.* ZNF469 frequently mutated in the brittle cornea syndrome (BCS) is a single exon gene possibly regulating the expression of several extracellular matrix components. Molecular genetics and metabolism. 2013;109(3):289-95.
170. Wu MC, Lee S, Cai T, *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics. 2011;89(1):82-93.
171. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics (Oxford, England). 2003;19(1):149-50.
172. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2009.
173. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017;45(D1):D158-d69.
174. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2', R package version 0.7.0 2016. Available from: https://CRAN.R-project.org/package=cowplot.
175. Siepel A, Bejerano G, Pedersen JS, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research. 2005;15(8):1034-50.
176. Siepel A, Haussler D. Phylogenetic hidden Markov models. In R. 2005. In: Statistical Methods in Molecular Evolution [Internet]. New York: Springer-Verlag; [325-51].
177. Kent WJ, Sugnet CW, Furey TS, *et al.* The human genome browser at UCSC. Genome research. 2002;12(6):996-1006.
178. Godefrooij DA, de Wit GA, Uiterwaal CS, *et al.* Age-specific Incidence and Prevalence of Keratoconus: A Nationwide Registration Study. American journal of ophthalmology. 2017;175:169-72.
179. Hayashi T, Huang J, Deeb SS. RINX(VSX1), a novel homeobox gene expressed in the inner nuclear layer of the adult retina. Genomics. 2000;67(2):128-39.
180. Semina EV, Mintz-Hittner HA, Murray JC. Isolation and characterization of a novel human paired-like homeodomain-containing transcription factor gene, VSX1, expressed in ocular tissues. Genomics. 2000;63(2):289-93.
181. Heon E, Mathers WD, Alward WL, *et al.* Linkage of posterior polymorphous corneal dystrophy to 20q11. Human molecular genetics. 1995;4(3):485-8.
182. Bechara SJ, Grossniklaus HE, Waring GO, 3rd, *et al.* Keratoconus associated with posterior polymorphous dystrophy. American journal of ophthalmology. 1991;112(6):729-31.
183. Blair SD, Seabrooks D, Shields WJ, *et al.* Bilateral progressive essential iris atrophy and keratoconus with coincident features of posterior polymorphous dystrophy: a case report and proposed pathogenesis. Cornea. 1992;11(3):255-61.
184. Driver PJ, Reed JW, Davis RM. Familial cases of keratoconus associated with posterior polymorphous dystrophy. American journal of ophthalmology. 1994;118(2):256-7.
185. Gasset AR, Zimmerman TJ. Posterior polymorphous dystrophy associated with keratoconus. American journal of ophthalmology. 1974;78(3):535-7.

186. Weissman BA, Ehrlich M, Levenson JE, *et al.* Four cases of keratoconus and posterior polymorphous corneal dystrophy. Optometry and vision science : official publication of the American Academy of Optometry. 1989;66(4):243-6.
187. Abu A, Frydman M, Marek D, *et al.* Deleterious mutations in the Zinc-Finger 469 gene cause brittle cornea syndrome. American journal of human genetics. 2008;82(5):1217-22.
188. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-60.
189. Burdon KP, Fogarty RD, Shen W, *et al.* Genome-wide association study for sight-threatening diabetic retinopathy reveals association with genetic variation near the GRB2 gene. Diabetologia. 2015;58(10):2288-97.
190. Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.
191. The International HapMap Project. Nature. 2003;426(6968):789-96.
192. Anderson CA, Pettersson FH, Clarke GM, *et al.* Data quality control in genetic case-control association studies. Nature protocols. 2010;5(9):1564-73.
193. Price AL, Patterson NJ, Plenge RM, *et al.* Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904-9.
194. Plotly Technologies Inc. Collaborative data science. Montréal, QC, Canada: Plotly Technologies Inc.; 2015. Available from: https://plot.ly.
195. Seelow D, Schuelke M, Hildebrandt F, *et al.* HomozygosityMapper--an interactive approach to homozygosity mapping. Nucleic acids research. 2009;37(Web Server issue):W593-9.
196. Baron RV, Kollar C, Mukhopadhyay N, *et al.* Mega2: validated data-reformatting for linkage and association analyses. Source code for biology and medicine. 2014;9(1):26.
197. Abecasis GR, Cherny SS, Cookson WO, *et al.* Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics. 2002;30(1):97-101.
198. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England). 2010;26(6):841-2.
199. Huttlin EL, Bruckner RJ, Paulo JA, *et al.* Architecture of the human interactome defines protein communities and disease networks. Nature. 2017;545(7655):505-9. Epub 2017 May 17.
200. Hein MY, Hubner NC, Poser I, *et al.* A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell. 2015;163(3):712-23.
201. Paz S, Vilasco M, Arguello M, *et al.* Ubiquitin-regulated recruitment of IkappaB kinase epsilon to the MAVS interferon signaling adapter. Mol Cell Biol. 2009;29(12):3401-12. Epub 2009 Apr 20.
202. Meylan E, Curran J, Hofmann K, *et al.* Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. Nature. 2005;437(7062):1167-72. Epub 2005 Sep 21.
203. Kumar H, Kawai T, Kato H, *et al.* Essential role of IPS-1 in innate immune responses against RNA viruses. J Exp Med. 2006;203(7):1795-803. Epub 2006 Jun 19.
204. Kawai T, Takahashi K, Sato S, *et al.* IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. Nat Immunol. 2005;6(10):981-8. Epub 2005 Aug 28.
205. Dong X, Feng P. Murine Gamma Herpesvirus 68 Hijacks MAVS and IKKβ to Abrogate NFκB Activation and Antiviral Cytokine Production. PLoS Pathog. 2011;7(11):e1002336. Epub 2011 Nov 10.
206. Kong XN, Yan HX, Chen L, *et al.* LPS-induced down-regulation of signal regulatory protein {alpha} contributes to innate immune activation in macrophages. J Exp Med. 2007;204(11):2719-31. Epub 007 Oct 22.
207. Marella M, Gaggioli C, Batoz M, *et al.* Pathological prion protein exposure switches on neuronal mitogen-activated protein kinase pathway resulting in microglia recruitment. The Journal of biological chemistry. 2004;280(2):1529-34. Epub 2004 Nov 4.
208. Drubay D, Gautheret D, Michiels S. A benchmark study of scoring methods for non-coding mutations. Bioinformatics (Oxford, England). 2018;34(10):1635-41.
209. Cervelli M, Amendola R, Polticelli F, *et al.* Spermine oxidase: ten years after. Amino acids. 2012;42(2):441-50.
210. Pegg AE. The function of spermine. IUBMB life. 2014;66(1):8-18.

211. Babbar N, Casero RA, Jr. Tumor necrosis factor-alpha increases reactive oxygen species by inducing spermine oxidase in human lung epithelial cells: a potential mechanism for inflammation-induced carcinogenesis. Cancer research. 2006;66(23):11125-30.
212. Chaturvedi R, de Sablet T, Coburn LA, *et al.* Arginine and polyamines in Helicobacter pylori-induced immune dysregulation and gastric carcinogenesis. Amino acids. 2012;42(2-3):627-40.
213. Xu H, Chaturvedi R, Cheng Y, *et al.* Spermine oxidation induced by Helicobacter pylori results in apoptosis and DNA damage: implications for gastric carcinogenesis. Cancer research. 2004;64(23):8521-5.
214. Sorkhabi R, Ghorbanihaghjo A, Taheri N, *et al.* Tear film inflammatory mediators in patients with keratoconus. International ophthalmology. 2015;35(4):467-72.
215. Lema I, Duran JA. Inflammatory molecules in the tears of patients with keratoconus. Ophthalmology. 2005;112(4):654-9.
216. Shetty R, Deshmukh R, Ghosh A, *et al.* Altered tear inflammatory profile in Indian keratoconus patients - The 2015 Col Rangachari Award paper. Indian journal of ophthalmology. 2017;65(11):1105-8.
217. Jun AS, Cope L, Speck C, *et al.* Subnormal cytokine profile in the tear fluid of keratoconus patients. PloS one. 2011;6(1):e16437.
218. Toprak I, Kucukatay V, Yildirim C, *et al.* Increased systemic oxidative stress in patients with keratoconus. Eye (London, England). 2014;28(3):285-9.
219. Kaldawy RM, Wagner J, Ching S, *et al.* Evidence of apoptotic cell death in keratoconus. Cornea. 2002;21(2):206-9.
220. Kim WJ, Rabinowitz YS, Meisler DM, *et al.* Keratocyte apoptosis associated with keratoconus. Experimental eye research. 1999;69(5):475-81.
221. Zhou Y, Koelling N, Fenwick AL, *et al.* Disruption of TWIST1 translation by 5' UTR variants in Saethre-Chotzen syndrome. Human mutation. 2018;39(10):1360-5.
222. Hornig NC, de Beaufort C, Denzer F, *et al.* A Recurrent Germline Mutation in the 5'UTR of the Androgen Receptor Causes Complete Androgen Insensitivity by Activating Aberrant uORF Translation. PloS one. 2016;11(4):e0154158.
223. Hsu AP, Zerbe CS, Foruraghi L, *et al.* IKBKG (NEMO) 5' Untranslated Splice Mutations Lead to Severe, Chronic Disseminated Mycobacterial Infections. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2018;67(3):456-9.
224. Bolze A, Boisson B, Bosch B, *et al.* Incomplete penetrance for isolated congenital asplenia in humans with mutations in translated and untranslated RPSA exons. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(34):E8007-e16.
225. Stallings RL, Doggett NA, Okumura K, *et al.* Chromosome 16-specific repetitive DNA sequences that map to chromosomal regions known to undergo breakage/rearrangement in leukemia cells. Genomics. 1992;13(2):332-8.
226. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2011;13:36.
227. Gusella JF, Wexler NS, Conneally PM, *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983;306(5940):234-8.
228. Verkerk AJ, Pieretti M, Sutcliffe JS, *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991;65(5):905-14.
229. Kremer EJ, Pritchard M, Lynch M, *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science (New York, NY). 1991;252(5013):1711-4.
230. Vafiadis P, Bennett ST, Todd JA, *et al.* Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. Nature genetics. 1997;15(3):289-92.
231. Pugliese A, Zeller M, Fernandez A, Jr., *et al.* The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. Nature genetics. 1997;15(3):293-7.
232. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007;447:932.
233. Dashnow H, Lek M, Phipson B, *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome biology. 2018;19(1):121.

234. Chaisson MJ, Huddleston J, Dennis MY, *et al.* Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517(7536):608-11.
235. Belkadi A, Bolze A, Itan Y, *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(17):5473-8.
236. Lelieveld SH, Spielmann M, Mundlos S, *et al.* Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. Human mutation. 2015;36(8):815-22.
237. John B, Lewis KR. Chromosome variability and geographic distribution in insects. Science (New York, NY). 1966;152(3723):711-21.
238. Gottesman, II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. The American journal of psychiatry. 2003;160(4):636-45.
239. Gottesman, II, Shields J. Genetic theorizing and schizophrenia. The British journal of psychiatry : the journal of mental science. 1973;122(566):15-30.
240. Gershon ES, Goldin LR. Clinical methods in psychiatric genetics. I. Robustness of genetic marker investigative strategies. Acta psychiatrica Scandinavica. 1986;74(2):113-8.
241. Leboyer M, Bellivier F, Nosten-Bertrand M, *et al.* Psychiatric genetics: search for phenotypes. Trends in neurosciences. 1998;21(3):102-5.
242. Zheng Y, Ge J, Huang G, *et al.* Heritability of central corneal thickness in Chinese: the Guangzhou Twin Eye Study. Investigative ophthalmology & visual science. 2008;49(10):4303-7.
243. Toh T, Liew SH, MacKinnon JR, *et al.* Central corneal thickness is highly heritable: the twin eye studies. Investigative ophthalmology & visual science. 2005;46(10):3718-22.
244. Landers JA, Hewitt AW, Dimasi DP, *et al.* Heritability of central corneal thickness in nuclear families. Investigative ophthalmology & visual science. 2009;50(9):4087-90.
245. Alsbirk PH. Corneal thickness. II. Environmental and genetic factors. Acta Ophthalmol (Copenh). 1978;56(1):105-13.
246. Colin J, Sale Y, Malet F, *et al.* Inferior steepening is associated with thinning of the inferotemporal cornea. Journal of refractive surgery (Thorofare, NJ : 1995). 1996;12(6):697-9.
247. Steele TM, Fabinyi DC, Couper TA, *et al.* Prevalence of Orbscan II corneal abnormalities in relatives of patients with keratoconus. Clin Exp Ophthalmol. 2008;36(9):824-30.
248. Cuellar-Partida G, Springelkamp H, Lucas SE, *et al.* WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. Human molecular genetics. 2015;24(17):5060-8.
249. Iglesias AI, Mishra A, Vitart V, *et al.* Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. Nat Commun. 2018;9(1):1864.
250. Auton A, Brooks LD, Durbin RM, *et al.* A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
251. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nature methods. 2011;9(2):179-81.
252. Das S, Forer L, Schonherr S, *et al.* Next-generation genotype imputation service and methods. Nature genetics. 2016;48(10):1284-7.
253. McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. Nature genetics. 2016;48(10):1279-83.
254. Fritsche LG, Igl W, Bailey JN, *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature genetics. 2016;48(2):134-43.
255. Loh PR, Danecek P, Palamara PF, *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. Nature genetics. 2016;48(11):1443-8.
256. Pruim RJ, Welch RP, Sanna S, *et al.* LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics (Oxford, England). 2010;26(18):2336-7.
257. Barrett JC, Fry B, Maller J, *et al.* Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics (Oxford, England). 2005;21(2):263-5.

258. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org1996-2010.
259. Frankish A, Diekhans M, Ferreira AM, *et al.* GENCODE reference annotation for the human and mouse genomes. Nucleic acids research. 2018.
260. Schaefer L, Iozzo RV. Biological functions of the small leucine-rich proteoglycans: from genetics to signal transduction. The Journal of biological chemistry. 2008;283(31):21305-9.
261. Michelacci YM. Collagens and proteoglycans of the corneal extracellular matrix. Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas. 2003;36(8):1037-46.
262. Liu CY, Birk DE, Hassell JR, *et al.* Keratocan-deficient mice display alterations in corneal structure. The Journal of biological chemistry. 2003;278(24):21672-7.
263. Yamaguchi Y, Mann DM, Ruoslahti E. Negative regulation of transforming growth factor-beta by the proteoglycan decorin. Nature. 1990;346(6281):281-4.
264. Schonherr E, Levkau B, Schaefer L, *et al.* Decorin-mediated signal transduction in endothelial cells. Involvement of Akt/protein kinase B in up-regulation of p21(WAF1/CIP1) but not p27(KIP1). The Journal of biological chemistry. 2001;276(44):40687-92.
265. Jarvelainen H, Puolakkainen P, Pakkanen S, *et al.* A role for decorin in cutaneous wound healing and angiogenesis. Wound repair and regeneration : official publication of the Wound Healing Society [and] the European Tissue Repair Society. 2006;14(4):443-52.
266. Mauviel A, Santra M, Chen YQ, *et al.* Transcriptional regulation of decorin gene expression. Induction by quiescence and repression by tumor necrosis factor-alpha. The Journal of biological chemistry. 1995;270(19):11692-700.
267. Mauviel A, Korang K, Santra M, *et al.* Identification of a bimodal regulatory element encompassing a canonical AP-1 binding site in the proximal promoter region of the human decorin gene. The Journal of biological chemistry. 1996;271(40):24824-9.
268. Wegrowski Y, Paltot V, Gillery P, *et al.* Stimulation of sulphated glycosaminoglycan and decorin production in adult dermal fibroblasts by recombinant human interleukin-4. The Biochemical journal. 1995;307 (Pt 3):673-8.
269. Vij N, Roberts L, Joyce S, *et al.* Lumican suppresses cell proliferation and aids Fas-Fas ligand mediated apoptosis: implications in the cornea. Experimental eye research. 2004;78(5):957-71.
270. Vuillermoz B, Khoruzhenko A, D'Onofrio MF, *et al.* The small leucine-rich proteoglycan lumican inhibits melanoma progression. Experimental cell research. 2004;296(2):294-306.
271. Liu CY, Kao WW. Lumican promotes corneal epithelial wound healing. Methods in molecular biology (Clifton, NJ). 2012;836:285-90.
272. Hayashi Y, Call MK, Chikama T, *et al.* Lumican is required for neutrophil extravasation following corneal injury and wound healing. Journal of cell science. 2010;123(Pt 17):2987-95.
273. Albig AR, Roy TG, Becenti DJ, *et al.* Transcriptome analysis of endothelial cell gene expression induced by growth on matrigel matrices: identification and characterization of MAGP-2 and lumican as novel regulators of angiogenesis. Angiogenesis. 2007;10(3):197-216.
274. Carlson EC, Liu CY, Chikama T, *et al.* Keratocan, a cornea-specific keratan sulfate proteoglycan, is regulated by lumican. The Journal of biological chemistry. 2005;280(27):25541-7.
275. Igwe JC, Gao Q, Kizivat T, *et al.* Keratocan is expressed by osteoblasts and can modulate osteogenic differentiation. Connective tissue research. 2011;52(5):401-7.
276. Bredrup C, Knappskog PM, Majewski J, *et al.* Congenital stromal dystrophy of the cornea caused by a mutation in the decorin gene. Investigative ophthalmology & visual science. 2005;46(2):420-6.
277. Wang IJ, Chiang TH, Shih YF, *et al.* The association of single nucleotide polymorphisms in the 5'-regulatory region of the lumican gene with susceptibility to high myopia in Taiwan. Molecular vision. 2006;12:852-7.
278. Majava M, Bishop PN, Hagg P, *et al.* Novel mutations in the small leucine-rich repeat protein/proteoglycan (SLRP) genes in high myopia. Human mutation. 2007;28(4):336-44.
279. Chen ZT, Wang IJ, Shih YF, *et al.* The association of haplotype at the lumican gene with high myopia susceptibility in Taiwanese patients. Ophthalmology. 2009;116(10):1920-7.
280. Zhang F, Zhu T, Zhou Z, *et al.* Association of lumican gene with susceptibility to pathological myopia in the northern han ethnic chinese. J Ophthalmol. 2009;2009:514306.

281. Lin HJ, Kung YJ, Lin YJ, *et al.* Association of the lumican gene functional 3'-UTR polymorphism with high myopia. Investigative ophthalmology & visual science. 2010;51(1):96-102.
282. Lin HJ, Wan L, Tsai Y, *et al.* The association between lumican gene polymorphisms and high myopia. Eye (London, England). 2010;24(6):1093-101.
283. Feng YF, Zhang YL, Zha Y, *et al.* Association of Lumican gene polymorphism with high myopia: a meta-analysis. Optometry and vision science : official publication of the American Academy of Optometry. 2013;90(11):1321-6.
284. Liao X, Yang XB, Liao M, *et al.* Association between lumican gene -1554 T/C polymorphism and high myopia in Asian population: a meta-analysis. International journal of ophthalmology. 2013;6(5):696-701.
285. Deng ZJ, Shi KQ, Song YJ, *et al.* Association between a lumican promoter polymorphism and high myopia in the Chinese population: a meta-analysis of case-control studies. Ophthalmologica Journal international d'ophtalmologie International journal of ophthalmology Zeitschrift fur Augenheilkunde. 2014;232(2):110-7.
286. He M, Wang W, Ragoonundun D, *et al.* Meta-analysis of the association between lumican gene polymorphisms and susceptibility to high Myopia. PloS one. 2014;9(6):e98748.
287. Wang GF, Ji QS, Qi B, *et al.* The association of lumican polymorphisms and high myopia in a Southern Chinese population. International journal of ophthalmology. 2017;10(10):1516-20.
288. Wang P, Li S, Xiao X, *et al.* High myopia is not associated with the SNPs in the TGIF, lumican, TGFB1, and HGF genes. Investigative ophthalmology & visual science. 2009;50(4):1546-51.
289. Yip SP, Leung KH, Ng PW, *et al.* Evaluation of proteoglycan gene polymorphisms as risk factors in the genetic susceptibility to high myopia. Investigative ophthalmology & visual science. 2011;52(9):6396-403.
290. Dai L, Li Y, Du CY, *et al.* Ten SNPs of PAX6, Lumican, and MYOC genes are not associated with high myopia in Han Chinese. Ophthalmic genetics. 2012;33(3):171-8.
291. Park SH, Mok J, Joo CK. Absence of an association between lumican promoter variants and high myopia in the Korean population. Ophthalmic genetics. 2013;34(1-2):43-7.
292. Li M, Zhai L, Zeng S, *et al.* Lack of association between LUM rs3759223 polymorphism and high myopia. Optometry and vision science : official publication of the American Academy of Optometry. 2014;91(7):707-12.
293. Okui S, Meguro A, Takeuchi M, *et al.* Analysis of the association between the LUM rs3759223 variant and high myopia in a Japanese population. Clinical ophthalmology (Auckland, NZ). 2016;10:2157-63.
294. Diskin S, Kumar J, Cao Z, *et al.* Detection of differentially expressed glycogenes in trabecular meshwork of eyes with primary open-angle glaucoma. Investigative ophthalmology & visual science. 2006;47(4):1491-9.
295. Pellegata NS, Dieguez-Lucena JL, Joensuu T, *et al.* Mutations in KERA, encoding keratocan, cause cornea plana. Nature genetics. 2000;25(1):91-5.
296. Akama TO, Nishida K, Nakayama J, *et al.* Macular corneal dystrophy type I and type II are caused by distinct mutations in a new sulphotransferase gene. Nature genetics. 2000;26:237.
297. Wentz-Hunter K, Cheng EL, Ueda J, *et al.* Keratocan expression is increased in the stroma of keratoconus corneas. Molecular medicine (Cambridge, Mass). 2001;7(7):470-7.
298. Birk DE. Type V collagen: heterotypic type I/V collagen interactions in the regulation of fibril assembly. Micron (Oxford, England : 1993). 2001;32(3):223-37.
299. Li X, Bykhovskaya Y, Canedo AL, *et al.* Genetic association of COL5A1 variants in keratoconus patients suggests a complex connection between corneal thinning and keratoconus. Investigative ophthalmology & visual science. 2013;54(4):2696-704.
300. Rong SS, Ma STU, Yu XT, *et al.* Genetic associations for keratoconus: a systematic review and meta-analysis. Scientific reports. 2017;7(1):4620.
301. Abu-Amero KK, Helwa I, Al-Muammar A, *et al.* Case-control association between CCT-associated variants and keratoconus in a Saudi Arabian population. Journal of negative results in biomedicine. 2015;14:10.
302. Liskova P, Dudakova L, Krepelova A, *et al.* Replication of SNP associations with keratoconus in a Czech cohort. PloS one. 2017;12(2):e0172365.

303. Li L, Li Y, Browning SR, *et al.* Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. PloS one. 2011;6(9):e24945.
304. Dickson SP, Wang K, Krantz I, *et al.* Rare variants create synthetic genome-wide associations. PLoS biology. 2010;8(1):e1000294.
305. Frazer KA, Murray SS, Schork NJ, *et al.* Human genetic variation and its contribution to complex traits. Nature reviews Genetics. 2009;10(4):241-51.
306. Maurano MT, Humbert R, Rynes E, *et al.* Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, NY). 2012;337(6099):1190-5.
307. Walters RG, Coin LJ, Ruokonen A, *et al.* Rare genomic structural variants in complex disease: lessons from the replication of associations with obesity. PloS one. 2013;8(3):e58048.
308. Walters RG, Jacquemont S, Valsesia A, *et al.* A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature. 2010;463(7281):671-5.
309. Schilter KF, Reis LM, Schneider A, *et al.* Whole-genome copy number variation analysis in anophthalmia and microphthalmia. Clinical genetics. 2013;84(5):473-81.
310. Siggs OM, Javadiyan S, Sharma S, *et al.* Partial duplication of the CRYBB1-CRYBA4 locus is associated with autosomal dominant congenital cataract. European journal of human genetics : EJHG. 2017;25(6):711-8.
311. Naderan M, Shoar S, Rezagholizadeh F, *et al.* Characteristics and associations of keratoconus patients. Contact lens & anterior eye : the journal of the British Contact Lens Association. 2015;38(3):199-205.
312. Gordon MO, Beiser JA, Brandt JD, *et al.* The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. Archives of ophthalmology. 2002;120(6):714-20; discussion 829-30.
313. Herndon LW, Weizer JS, Stinnett SS. Central corneal thickness as a risk factor for advanced glaucoma damage. Archives of ophthalmology. 2004;122(1):17-21.
314. Kim JW, Chen PP. Central corneal pachymetry and visual field progression in patients with open-angle glaucoma. Ophthalmology. 2004;111(11):2126-32.
315. Charlesworth J, Kramer PL, Dyer T, *et al.* The path to open-angle glaucoma gene discovery: endophenotypic status of intraocular pressure, cup-to-disc ratio, and central corneal thickness. Investigative ophthalmology & visual science. 2010;51(7):3509-14.
316. Chaerkady R, Shao H, Scott SG, *et al.* The keratoconus corneal proteome: loss of epithelial integrity and stromal degeneration. Journal of proteomics. 2013;87:122-31.

Appendix 1 – An example of the command for adding confidence tags to genotypes in a VCF file using the VariantFiltration tool from GATK.

```
## This script was developed by Dr. Blackburn
java -Xmx4g -jar /path/to/program/GenomeAnalysisTK.jar \
-R reference.fa \
-T VariantFiltration \
-o output_file.vcf \
--variant input_file.vcf \
--genotypeFilterExpression "DP >= 10 && GQ >= 20" \
--genotypeFilterName "High_Confidence" \
--genotypeFilterExpression "DP < 10 && GQ < 20" \
--genotypeFilterName "Low_Confidence" \
--genotypeFilterExpression "DP < 10 && GQ >= 20" \
--genotypeFilterName "Low_coverage_High_quality" \
--genotypeFilterExpression "DP >= 10 && GQ < 20" \
--genotypeFilterName "High_coverage_Low_quality"
```

Appendix 2 – R script for converting genotypes with low coverage or low quality scores to missing.

This method relies on the addition of confidence tags to the VCF as described in Appendix 1.

```
## This script was developed with help from Dr. Blackburn and Dr. McComish
# Load the VCF (with confidence tags added to the genotypes)
geno <- read.table("input_file.txt", sep = "\t", header = T, as.is=TRUE,
                   na.strings = ".")

# Subset to just the genotype fields of the VCF
geno_only <- geno[,10:ncol(geno)]

#Converts all low coverage/quality/confidence calls to "./." (missing) based on
#the VariantFiltration tool from GATK
geno_only < - as.data.frame(sapply(geno_only, gsub, pattern = "./.(:.*Low.*)",
                            replacement = "\\./\\.\\1"))

# Add back the other columns
geno_clean <- cbind2(geno[, 1:9], geno_only)

# Write table for re-merging with the VCF header
write.table(geno_clean, file = "output_file.txt", sep = "\t", row.names = F,
            col.names = F, quote = F, na = ".")

## to create a VCF file again, concatenate a copy of the header from the
input_file.vcf onto the output file.
```

Appendix 3 – Custom script used to annotate VCF files with ANNOVAR, including producing an output file containing a complete header line with the sample IDs present in the VCF file.

```
#To annotate VCF files with the 2017Jun01 version of ANNOVAR
#The start of the input filename (without the last two fields which for my files
#is the date of creation and the file extension) will be used as the output file
#name with "_Annotated_<date>" added at the end
## ie. if the input file is file_05082016.vcf, the output will be
#file_Annotated_07082016 (ANNOVAR produces a .vcf, .txt and .avinput files)
#The number of threads to use is indicated by the first argument
#Use a thread of 1 unless told otherwise (it can slow down the annotation if >1)
#The annotated TXT ANNOVAR output file does not contain the sample IDs.
#The output file with the suffix "_Annotated_with_Header_<date>.txt" has a
#complete header including sample IDs and is otherwise identical.

#Usage: bash ANNOVAR_2017_2.sh <number_of_threads> <filename_or_filenames>
module load annovar/2017Jun01
now=`date +"%d%m%Y"`

## Run ANNOVAR on one or more VCF files
for file in "${@:2}"
do
filename=`echo ${file} | cut -d "." -f 1`
filename_start=`echo ${filename} | rev | cut -d _ -f 2- | rev`
table_annovar.pl $file \
/gd/apps/annovar-2017Jun01/humandb -buildver hg19 \
-out ${filename_start}_Annotated2017Jun01_${now} \
-remove -otherinfo -thread ${1} \
-protocol refGene,avsnp147,popfreq_max_20150413,\
1000g2015aug_all,1000g2015aug_eur,1000g2015aug_afr,1000g2015aug_amr,\
1000g2015aug_eas,1000g2015aug_sas,esp6500siv2_all,esp6500siv2_ea,esp6500siv2_aa,\
exac03,exac03nontcga,exac03nonpsych,kaviar_20150923,hrcr1,gme,gnomad_exome,\
gnomad_genome,clinvar_20170130,cadd13,avsift,gerp++elem,dbnsfp31a_interpro,\
spidex,dbscsnv11,mitimpact24,dbnsfp33a,gwava,eigen,fathmm \
-operation g,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f,f \
-nastring . -vcfinput
# Add a header line to the annotated TXT file including the sample IDs
grep -w "#CHROM" ${file} | cut -f 2- > VCF_header.txt
head -1 ${filename_start}_Annotated2017Jun01_${now}.hg19_multianno.txt \
> ANNO_header.txt
paste -d'\t' ANNO_header.txt ~/Scripts/ANNOVAR_Additional_Column_Headings.txt \
VCF_header.txt > full_header.txt
tail -n+2 ${filename_start}_Annotated2017Jun01_${now}.hg19_multianno.txt | \
cat full_header.txt - > ${filename_start}_Annotated_with_Header_${now}.txt
#remove unnecessary intermediate files
rm VCF_header.tzt
rm ANNO_header.txt
rm full_header.txt
done
```

Appendix 4 – An example command for extracting regions from a variant caller format (VCF) file using BCFtools.

```
bcftools view -R region_file.txt -S Sample_List.txt -c 1 -O v -o output_file.vcf \
input_file.vcf.gz
```

Appendix 5 – R code for plotting the first two principle components from the principle components analysis (PCA) using data from the keratoconus families and HapMap Phase III.

```
#read in the data
PCA <- read.table("Separate_Family_Colours_merge-data.hapmap3r2.pruned.pca.evec",
                  header = FALSE ,skip=1)

# Add column names
colnames(PCA) <- c("Sample ID", "PC1", "PC2", "Population")

# Generate scatter plot
library("ggplot2")
ggplot(PCA, aes(x = PC1, y = PC2, color=Population)) +
       geom_point() + theme_bw() +
       scale_x_continuous(limits = c(-0.1,0.05)) +
       scale_y_continuous(limits = c(-0.05,0.1)) +
       scale_colour_brewer(palette = "Dark2", direction = -1)
```

Appendix 6 – Example R code for generating 3D plots to visualise IBD estimates using plotly in R.

```
#Load the plotly package
library(plotly)

#Link to a plotly account
Sys.setenv("plotly_username"="your_username")
Sys.setenv("plotly_api_key"="your_api_key")

# Read in the file
IBD <- read.table(file="plink_IBD_estimates.txt", header=T, sep = "\t")

# Generate a 3D plot of Z0 vs Z1 vs Z2
p <- plot_ly(IBD, x= ~Z0, y = ~Z1, z = ~Z2, color = ~Relationship, colors =
c("#0072B2","#D55E00", "#CC79A7", "#F0E442", "#009E73"), hoverinfo = 'text',
text = ~paste(IID1, ' vs ', IID2, '(PI_HAT = ',PI_HAT,')')) %>% add_markers() %>%
layout(legend = list(x= 0.8, y = 0.8, z = 0.01), title = "Plot title", scene =
list(xaxis = list(title = "Plink Z0", range = c(0,1), minorgridcount = 5,
minorgridwidth = 2), yaxis = list(title = "Plink Z1", range = c(0,1),
minorgridcount = 5, minorgridwidth = 2), zaxis = list(title = "Plink Z2", range =
c(0,1), minorgridcount = 5, minorgridwidth = 2)))
```

```
#Create online interactive plot
chart_link = api_create(p, filename="output_filename")
```

Appendix 7 – R code for plotting homozygosity scores for all autosomes in a single plot.

```
# Read in the file downloaded from HomozygosityMapper
HomScores <- read.table(file = "homozygosity_scores.txt", sep = "\t", header = T)

## Plot all autosomes in a single plot
library(ggplot2)

ggplot(data=HomScores, aes(x=position, y=score)) +
  geom_bar(colour = "black", stat="identity") +
  labs(y = "Homozygosity Score", x ="Chromosome") +
  scale_y_continuous(limits = c(0,350), breaks = seq(0, 350, 50), expand = c(0,0))
+
  facet_grid (~ chromosome, scales = "free_x", space = "free_x", switch = "x") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
        panel.grid.minor = element_blank(), strip.background = element_blank(),
        panel.grid.major.x = element_blank(), panel.spacing = unit(0, "lines"),
        panel.border = element_rect(size = 0.25, color = "grey80")) + theme_bw()
```

Appendix 8 – The PLINK command for identifying runs of homozygosity.

The '--homozyg group' option was used to generate a report of homozygous regions that overlap in at least two individuals. This command produces a file with the name convention '<familyID>.hom.overlap'.

```
plink --bfile ~/path/to/family-specific/independent_SNPs --homozyg group --out
<familyID>
```

Appendix 9 – Method for remove duplicate cM positions from Merlin format Map files for linkage analysis.

Prior to reading the Merlin format map files into R, a sed command was used to convert any white-space to tabs to ensure the files were tab delimitated:

```
sed -i 's/ \+ /\t/g' merlin_map.01
```

The following R commands, developed by Johanna Jones and Michael Sumner, were then applied to the Merlin format map files to identify and alter duplicate cM positions.

```
# Load in the map file
map <- read.table("merlin_map.01", sep = "\t", header = T,
                  as.is=TRUE,na.strings=".")
```

```
## To see if there are duplicate positions
# a “TRUE” result means there's no duplicates
# a "FALSE" result means there's duplicates in the positions column
length(unique(map$POSITION)) == nrow(map)

#load the dplyr package
library(dplyr)

# To add a very small value to the second SNP in the pair with duplicate values so
they are # no longer duplicates
input_order <- unique(as.character(map$POSITION))
altered <- bind_rows(lapply(split(map, map$POSITION),
                                function(x) {
                                  if (nrow(x) > 1) {
                                       x$POSITION <- x$POSITION +
                                       cumsum(c(0,rep(0.0000001,nrow(x) - 1)))
                                       }
                                       x
                                       })[input_order])

#Check it worked
length(unique(altered$POSITION)) == nrow(altered)
# If it worked the output will be "TRUE"

# Write the table to a file (saving over the old file)
write.table(altered, file = "merlin_map.17", sep = "\t", row.names = F,
            col.names = T, quote = F)
```

Appendix 10 – Haplotype estimation based on the most likely pattern of gene flow using MERLIN.

Ensure genotype errors are removed in MERLIN prior to this analysis.

```
merlin -d <chr_of_interest>_wiped.dat -p <chr_of_interest>_wiped.ped -m
merlin_map.<chr_of_interest_number> --best
```

Appendix 11 – Example commands for determining the mean depth and standard deviation across a region in multiple individuals and extracting a file containing regions with a mean depth below 10.

This was a three step process. The first step used SAMTools to extract the region of interest from the aligned BAM files for the relevant individuals using the following command; where the region of interest is in the format ‘chr#:1234-2345’ and bam.txt contains a list of the aligned BAM files and the path for each included individual in the format ‘/path/to/bam/finename.bam’:

```
samtools depth -a -r <region_of_interest> -f bams.txt output_filename.txt
```

The second step used R to calculate the mean depth and standard deviation across the region and all individuals. Bases with a read depth below 10 are then written to a BED format file.

```
# Read in the annotated variants file
depth <- read.table(file = "output_filename.txt",
                    sep = "\t", header = FALSE, as.is = TRUE, na.strings = ".",
quote = "")

## Create column headings, using the list of BAM files 'bams.txt' as above
# Read in sample IDS
bams <- read.table(file = " bams.txt",
                   sep = "/", header = FALSE, as.is = TRUE, na.strings = ".",
quote = "")

# Extract the file names (what ever is present before the "." and the extension)
samples <- gsub("\\..*", "", bams[,8])

# Add column names
colnames(depth) <- c("CHR", "POS", unlist(samples, use.names = FALSE))

## Calculate the mean depth and standard deviation for each position
# make sure to change the relevant columns as required (depending on the number of
people)
depth$depth_mean <- rowMeans(depth[,3:7])
depth$depth_sd <- apply(depth[,3:7], 1, sd)

## Calculating the mean and sd of the means (across all samples and bases in the
region)
mean(depth$depth_mean)
sd(depth$depth_mean)

## Extracting Regions with less than a depth of 10 ##
poor_cov <- depth[(depth$depth_mean<10),]

#Create a BED file
# Make sure to use the right columns for the mean depth (in particular)
poor_cov_bed <- poor_cov[,c(1:2,2,8)]
colnames(poor_cov_bed) <- c("CHR", "START", "END", "MEAN_DEPTH")

# write out into a BED file (without headers):
write.table(poor_cov_bed, file = "Regions_with_depth_below_10.bed",
            sep = "\t", row.names = F, col.names = F, quote = F, na = ".")
```

The final step uses BEDTools to merge consecutive positions with a mean depth below 10 into single regions. An example of the command is as follows:

```
bedtools merge -i Regions_with_depth_below_10.bed >
Regions_with_depth_below_10_ranges.bed
```

Appendix 12 – Custom script for detecting and removing genotyping errors using Pedwipe and conducting parametric linkage analysis using MERLIN.

```
# Usage: bash merlin_linkage_autosomes.sh | tee -a Log.txt
# Load MERLIN
module load merlin

# Create a log file
date > Log.txt

# Do the following for each autosome
for chr in {01..22}
do

if [[ -f merlin_map.${chr} && -f merlin_ped.${chr} && -f merlin_data.${chr} && \
      -f merlin_model ]]

then
    echo "starting chr$chr"
    echo "chr${chr}_wiped.dat and chr${chr}_wiped.ped are absent"
    echo "MERLIN input files merlin_map.${chr} , merlin_ped.${chr} , \
          merlin_data.${chr} and merlin_model are present"

   # Run PEDSTATS
    echo "Running PEDSTATS for chr$chr"
    pedstats -d merlin_data.${chr} -p merlin_ped.${chr}
    mv pedstats.markerinfo chr${chr}_pedstats.markerinfo

    # Error identification and make a copy of the data labelled by chromosome in
    #the merlin.err file
    echo "Running an error check for chr${chr}"
    merlin -d merlin_data.${chr} -p merlin_ped.${chr} -m merlin_map.${chr} --error
    cp merlin.err chr${chr}_merlin.err

    echo "merlin.err has been renamed chr${chr}_merlin.err"

    # Fixing errors with PEDWIPE
    echo "Removing problematic SNPs with PEDWIPE for chr${chr}"
    pedwipe -d merlin_data.${chr} -p merlin_ped.${chr}
    # This uses merlin.err (the output of the error identification), which will be
    #re-written for each chromosome, but I have copied this for each chromosome in
    #a file starting with "chr", followed by the chromosome number

    # Remove the generic named error file
    rm merlin.err

    # Rename output PEDWIPE files
    mv wiped.dat chr${chr}_wiped.dat
    mv wiped.freq chr${chr}_wiped.freq
    mv wiped.ped chr${chr}_wiped.ped
```

```
    echo "wiped.dat has been renamed chr${chr}_wiped.dat"
    echo "wiped.freq has been renamed chr${chr}_wiped.freq"
    echo "wiped.ped has been renamed chr${chr}_wiped.ped"

    # Run parametric linkage analysis with MERLIN
    echo "Running parametric linkage analysis with MERLIN for chr$chr"
    merlin -d chr${chr}_wiped.dat -p chr${chr}_wiped.ped -m merlin_map.$chr \
    --model merlin_model --tabulate --markerNames --pdf --quiet
    mv merlin.pdf chr${chr}_merlin.pdf
    mv merlin-parametric.tbl chr${chr}_merlin-parametric.tbl

    echo "merlin.pdf has been renamed chr${chr}_merlin.pdf"
    echo "merlin-parametric.tbl has been renamed chr${chr}merlin-parametric.tbl"
    echo "Done with chr$chr"

else
    echo "starting chr$chr"
    echo "WARNING - At least one of the required input files is absent"
    echo "Skipping chr$chr"
fi
done

# Removing unnecessary clutter from the log file
sed -i '/Right Conditional:/d' Log.txt
sed -i '/Singlepoint:/d' Log.txt
sed -i '/Scanning Chromosome:/d' Log.txt
```

Appendix 13 – R code for plotting parametric linkage results for all autosomes in a single plot.

Merlin outputs separate files for each chromosome. Prior to plotting, these files need to be concatenated together.

```
## Loading in my data
LODs <- read.table("Merlin_output_all_autosomes_concatenated.txt", header = F, sep
= "\t")

#Add column names
colnames(LODs) <- c("CHR","POS","LOD")
## NB the POS is in Morgans

## Convert Morgans to cMs
LODs$POS <- LODs$POS * 100

library("ggplot2")

# Create the plot
ggplot(data = LODs, aes(x = POS, y = LOD)) + geom_line(size = .75) +
```

```
labs(y = "LOD score", x ="Chromosome") + geom_hline(yintercept = -2, colour =
"red") +
  geom_hline(yintercept = 0, linetype = 5) +
  facet_grid (~ CHR, scales = "free_x", space = "free_x", switch = "x")
  + theme_bw() +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
        panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        strip.background = element_blank(), panel.spacing = unit(0, "lines"),
        panel.border = element_rect(size = 0.25, color = "grey80"))
```

Appendix 14 – Identity-by-descent estimates for KSA197 and KCNSW01.

Each row outlines the data for a pair of individuals where the ID1 and ID2 columns contain the individual IDs; the EZ is the expected identity-by-descent (IBD) proportion based on the reported relationship; Z0, Z1 and Z2 are the probabilities that the individuals share zero, one or two alleles identity by descent at a given locus (respectively); and the PI_HAT represents the overall IBD proportion.

ID1	ID2	Reported Relationship	EZ	Z0	Z1	Z2	PI_HAT
KSA197.0	KSA197.1	Parent-Offspring	0.5	0.00	0.62	0.38	0.69
KSA197.0	KSA197.2	Full Sibling	0.5	0.12	0.32	0.56	0.72
KSA197.0	KSA197.3	Parent-Offspring	0.5	0.00	0.62	0.38	0.69
KSA197.0	KSA197.4	Parent-Offspring	0.5	0.00	0.63	0.37	0.68
KSA197.1	KSA197.2	Parent-Offspring	0.5	0.00	0.62	0.38	0.69
KSA197.1	KSA197.3	Second Cousins	0.03	0.58	0.05	0.37	0.40
KSA197.1	KSA197.4	Grandparent-Grandchild	0.25	0.27	0.36	0.37	0.55
KSA197.2	KSA197.3	Parent-Offspring	0.5	0.00	0.61	0.39	0.70
KSA197.2	KSA197.4	Avuncular Pair	0.25	0.27	0.36	0.37	0.55
KSA197.3	KSA197.4	Grandparent-Grandchild	0.25	0.33	0.31	0.37	0.52
KCNSW01-1	KCNSW01-2	Parent-Offspring	0.5	0.01	0.79	0.20	0.60
KCNSW01-1	KCNSW01-3	Full Sibling	0.5	0.20	0.40	0.41	0.60
KCNSW01-1	KCNSW01-4	Parent-Offspring	0.5	0.00	0.81	0.19	0.59
KCNSW01-1	KCNSW01-5	Parent-Offspring	0.5	0.00	0.81	0.19	0.60
KCNSW01-1	KCNSW01-6	Full Sibling	0.5	0.20	0.45	0.34	0.57
KCNSW01-1	KCNSW01-7	Full Sibling	0.5	0.23	0.43	0.34	0.55
KCNSW01-1	KCNSW01-8	Full Sibling	0.5	0.22	0.40	0.38	0.58
KCNSW01-1	KCNSW01-9	Full Sibling	0.5	0.22	0.37	0.41	0.60
KCNSW01-1	KCNSW01-10	Full Sibling	0.5	0.25	0.40	0.35	0.55
KCNSW01-1	KCNSW01-11	Full Sibling	0.5	0.22	0.40	0.38	0.58
KCNSW01-2	KCNSW01-3	Avuncular Pair	0.25	0.39	0.41	0.20	0.40
KCNSW01-2	KCNSW01-4	Grandparent-Grandchild	0.25	0.41	0.39	0.20	0.39
KCNSW01-2	KCNSW01-5	Grandparent-Grandchild	0.25	0.39	0.41	0.20	0.41

ID1	ID2	Reported Relationship	EZ	Z0	Z1	Z2	PI_HAT
KCNSW01-2	KCNSW01-6	Avuncular Pair	0.25	0.43	0.37	0.20	0.38
KCNSW01-2	KCNSW01-7	Avuncular Pair	0.25	0.45	0.36	0.19	0.37
KCNSW01-2	KCNSW01-8	Avuncular Pair	0.25	0.41	0.40	0.20	0.40
KCNSW01-2	KCNSW01-9	Avuncular Pair	0.25	0.34	0.47	0.19	0.43
KCNSW01-2	KCNSW01-10	Avuncular Pair	0.25	0.40	0.40	0.20	0.40
KCNSW01-2	KCNSW01-11	Avuncular Pair	0.25	0.41	0.40	0.19	0.39
KCNSW01-3	KCNSW01-4	Parent-Offspring	0.5	0.00	0.82	0.18	0.59
KCNSW01-3	KCNSW01-5	Parent-Offspring	0.5	0.00	0.81	0.19	0.60
KCNSW01-3	KCNSW01-6	Full Sibling	0.5	0.15	0.42	0.43	0.64
KCNSW01-3	KCNSW01-7	Full Sibling	0.5	0.23	0.34	0.42	0.60
KCNSW01-3	KCNSW01-8	Full Sibling	0.5	0.16	0.40	0.44	0.64
KCNSW01-3	KCNSW01-9	Full Sibling	0.5	0.13	0.41	0.45	0.66
KCNSW01-3	KCNSW01-10	Full Sibling	0.5	0.17	0.35	0.48	0.65
KCNSW01-3	KCNSW01-11	Full Sibling	0.5	0.14	0.43	0.42	0.64
KCNSW01-4	KCNSW01-5	Unrelated	0	0.81	0.00	0.19	0.19
KCNSW01-4	KCNSW01-6	Parent-Offspring	0.5	0.00	0.81	0.19	0.60
KCNSW01-4	KCNSW01-7	Parent-Offspring	0.5	0.00	0.82	0.18	0.59
KCNSW01-4	KCNSW01-8	Parent-Offspring	0.5	0.00	0.82	0.18	0.59
KCNSW01-4	KCNSW01-9	Parent-Offspring	0.5	0.00	0.80	0.20	0.60
KCNSW01-4	KCNSW01-10	Parent-Offspring	0.5	0.00	0.80	0.20	0.60
KCNSW01-4	KCNSW01-11	Parent-Offspring	0.5	0.00	0.81	0.19	0.59
KCNSW01-5	KCNSW01-6	Parent-Offspring	0.5	0.00	0.81	0.19	0.60
KCNSW01-5	KCNSW01-7	Parent-Offspring	0.5	0.00	0.81	0.19	0.59
KCNSW01-5	KCNSW01-8	Parent-Offspring	0.5	0.00	0.80	0.20	0.60
KCNSW01-5	KCNSW01-9	Parent-Offspring	0.5	0.00	0.82	0.18	0.59
KCNSW01-5	KCNSW01-10	Parent-Offspring	0.5	0.00	0.81	0.19	0.59
KCNSW01-5	KCNSW01-11	Parent-Offspring	0.5	0.00	0.81	0.19	0.59
KCNSW01-6	KCNSW01-7	Full Sibling	0.5	0.25	0.41	0.34	0.54
KCNSW01-6	KCNSW01-8	Full Sibling	0.5	0.24	0.39	0.37	0.56
KCNSW01-6	KCNSW01-9	Full Sibling	0.5	0.23	0.39	0.37	0.57
KCNSW01-6	KCNSW01-10	Full Sibling	0.5	0.25	0.41	0.35	0.55
KCNSW01-6	KCNSW01-11	Full Sibling	0.5	0.13	0.41	0.45	0.66
KCNSW01-7	KCNSW01-8	Full Sibling	0.5	0.16	0.39	0.44	0.64
KCNSW01-7	KCNSW01-9	Full Sibling	0.5	0.21	0.35	0.44	0.61
KCNSW01-7	KCNSW01-10	Full Sibling	0.5	0.20	0.40	0.40	0.60
KCNSW01-7	KCNSW01-11	Full Sibling	0.5	0.22	0.40	0.38	0.58
KCNSW01-8	KCNSW01-9	Full Sibling	0.5	0.22	0.39	0.39	0.59
KCNSW01-8	KCNSW01-10	Full Sibling	0.5	0.18	0.41	0.42	0.62
KCNSW01-8	KCNSW01-11	Full Sibling	0.5	0.24	0.39	0.37	0.56
KCNSW01-9	KCNSW01-10	Full Sibling	0.5	0.15	0.39	0.46	0.65

ID1	ID2	Reported Relationship	EZ	Z0	Z1	Z2	PI_HAT
KCNSW01-9	KCNSW01-11	Full Sibling	0.5	0.16	0.40	0.44	0.64
KCNSW01-10	KCNSW01-11	Full Sibling	0.5	0.20	0.37	0.43	0.62

Appendix 15 – An example PLINK command for standard association analysis (chi squared tests).

```
plink --file <input_file_prefix> --assoc --ci 0.95 --mind 0.1 –geno 0.1 --hardy --
hwe 0.001 --out <output_file_prefix>
```

Appendix 16 – An example of the SAMtools command used to extract the depth information from multiple BAM files.

The file 'List_of_bams.txt' is a file containing the file names, and path to, all of the BAM files to be included.

```
samtools depth -a -r <chromosomal_region> -f List_of_bams.txt > \
<locus>_Coverage_All_Samples.txt
```

Appendix 17 – An example of the R script for determining mean depth and the standard deviation across all samples and plotting these data using ggplot2.

```
### Calculating variables for plotting ###
# Read in the file
depth <- read.table(file = "<locus>_Coverage_All_Samples.txt", sep = "\t", header
= FALSE, as.is = TRUE, na.strings = ".", quote = "")

## Adding column names
# Read in sample IDs
samples <- read.table(file = "Sample_IDs.txt", sep = "\t", header = FALSE, as.is =
TRUE, na.strings = ".", quote = "")

# Add column names
colnames(depth) <- c("CHR", "POS", unlist(samples, use.names = FALSE))

## Calculate values for all individuals
depth$depth_mean <- rowMeans(depth[,3:242])
depth$depth_sd <- apply(depth[,3:242], 1, sd)
depth$depth_meanMINUSsd <- depth$depth_mean - depth$depth_sd
depth$depth_meanPLUSsd <- depth$depth_mean + depth$depth_sd

# write out into a file:
write.table(depth, file = "<locus>_Coverage_All_Samples_with_Header.txt", sep =
"\t", row.names = F, col.names = T, quote = F, na = ".")
```

```
### Plotting mean and sd across the region ###
library("ggplot2")

ggplot(data=depth, aes(x = POS, y = depth_mean)) +
  geom_line() +
  geom_ribbon(aes(ymin= depth_meanMINUSsd, ymax = depth_meanPLUSsd), alpha = 0.3,
fill= "blue") +
  labs(y = "Depth", x ="Position on chromsome 9 (bp)") +
  scale_x_continuous(limits = c(<region_start-500bp>, (<region_end+500bp>),
  expand = c(0, 0)) + #the re-sequenced region +-500bp
  scale_y_continuous(limits = c(0,450), expand = c(0, 0)) +
  geom_hline(yintercept = 10, colour = "red") + theme_bw()
```